النماذج اللغوية الكبيرة العربية: تطوراتها، وتحدياتها وتطبيقاتها

بسم الله الرحمن الرحيم

1. مقدمة

لا شك أن الناظر في تطورات السنة الماضية لن تُخطأ عينهُ تألق وانتشار مجال معالجة اللغات الطبيعية مع كل المجالات سواء كانت بحثية أو تطبيقية وغالب هذه الطفرة كانت مع تطور النماذج اللغوية الضخمة فكان لزامًا أن يُكتب فيه ما يلم الشعث ويجمع الشتات بما يتعلق باللغة العربية فكانت هذه الفصول إذ حُصِرَ فيها غالب ما نُشر من نماذج عربية وتلخيص جوانبها البحثية. وقبلها بُدءَ بمقدمات مهمة تتعلق بالمصطلحات وتاريخ العلم وطُرُق تقييم النماذج .[1]
ذُكرت النماذج اللغوية على مجموعتين الأولى كانت في ذات الهيكلية المفككة ثم عُرّجَ بعدها على النوعين الآخرين (المرمزة والمرمزة المفككة) نظرًا إلى أن غالب الجهود من فترة انصبت على الهيكلية الأولى وكان لا بُدَّ من الإشارة للجميع فاختيرت هذه المنهجية .[2]ثم ذكرت أهم تطبيقات النماذج اللغوية وختاما ذُكرت توصيات مهمة تتعلق بهذا المجال [3].إضافات المسحيات

1.1. ما يدور عليه الفصل

الفصل يدور على هذه النقاط :

- المقدمات الأولية المهمة في المجال
- أساليب تقييم النماذج العربية
- حصر النماذج العربية
- أهم تطبيقاتها
- خاتمة وتوصيات.

2.1. أهم أسئلة الفصل

- ما هي الجهود المبذولة في تطوير النماذج العربية؟
- ما هي التطبيقات العملية التي تميزت بها هذه النماذج؟
- ما هي أساليب تقييم النماذج وأهم الجهود العربية فيها؟

---

[1]ولم نتعرض بشكل مفصل للمدونات النصية لأن هناك فصلا مفردًا عنها .

[2]يَحسُنُ بقارئ هذه الأوراق أن يكون على اطلاع يسير بهيكلية المحولات أجزاءها وسير تدريبها.

[3]يجدر الإشارة إلى أن نظام الإحالة اعتمد على طريقة chicago وقد استعملنا الأرقام للإحالة على المصادر فما كان بين قوسين [] فهو مصدر باللغة الانجليزية وما كان فيه حرف عين [ع] فهو مصدر باللغة العربية .

2. الباب الأول توطئة تاريخية واصطلاحية

سنتكلم في هذا الباب عن توطئة فيها أهم المصطلحات المهمة التي سيكثر استخدامها في هذه الأوراق مُعتمدين في ترجمتها على معجم سدايا للبيانات والذكاء الاصطناعي [ع6] ثم سنتكلم بشكل مختصر جدا عن تاريخ المجال وصولا لنماذج المحولات ونخلط فيه إشارات عن تاريخ مجال معالجة اللغات مع اللغة العربية وأبرز محطاتها، ولن نطيل فيه نظرًا إلى أن هناك فصلا كاملا عن تاريخ المجال ولكن نُلمح باختصار إلى أهم المحطات وخاصة ما يتعلق بهيكلية المحولات.

1.2. المصطلحات المُستخدمة

- المقسم Tokenizer وهو أداة تُستخدم لتقسيم النصوص المكتوبة بناءً على مجموعة أجزاء لغوية محدودة (مميزة) تم بناءها مُسبقًا
- الجزء اللغوي Token وهو قطعة نصية (قد تكون أقل من كلمة أو أكثر أو مساوية) تم استخراجها باستعمال أحد خوارزميات الضغط النصي لتكوين مجموعة محدودة من الأجزاء التي يُمكنُ تمثيل المدونات النصية بها .
- خوارزمية BPE وهي أحد أشهر الخوارزميات المُستعملة في ضغط النصوص ويكثر استعمالها في النماذج اللغوية. وهذا الاختصار (BPE) يعني Byte Pair Encoding أي ترميز أزواج البايت.
- هيكلية المحولات :Transformers وهي أحد أشهر وأحدث بُنى الشبكات العصبية الصادرة عام 2017 وأول ما صدرت كانت لمهمة الترجمة مكوَّنةً من جُزئين الأول مُرَمِّزْ (مشفر) والثاني مفكك الترميز .
- المحولات المرمِّزة :Encoder هيكلية تستعمل في تدريب النموذج جزءَ المرمز فقد وتُستعمل عادة في مهام الفهم اللغوي .
- المحولات المفككة :Decoder وهذه تستعمل فقط جزءَ المفكك من المحولات وتُستعمل عادة في توليد النصوص.
- المعامل :Parameter رقم متغير يتغير أثناء عملية تدريب نموذج الشبكة العصبية (بغض النظر عن بُنيتها) يتعلم بتغييره النموذج النمط المخفي بين المُدخلات والمُخرجات .
- المدونة :Corpus\Dataset مجموعة نصية وفي سياق النماذج اللغوية تكون خامًا غير مُوسَّمة تُستعمل في تعليم النموذج لغةً معينة .
- النموذج اللغوي الضخم :Large Language Model مُصطلح يشير إلى نموذج ذي عدد كبير من المعاملات وفي الفترة الأخيرة أُصطلِحَ على ما تجاوز مليار مُعامل.
- المحسِّن :Optimizer خوارزمية تُستخدم في تحديث معاملات النموذج وتغييرها أثناء عملية التدريب.
- التضمين :Embedding متجه رياضي ذي أرقام حقيقية يُستعمل في سياق النماذج اللغوية لتمثيل مركب نصي سواء كان كلمةً أو أكثر .
- خوارزمية :PPO أحد خوارزميات التعلم التعزيزي لتحسين أداء الوكلاء في المهمات الصعبة (وفي حالة النماذج اللغوية أُستعملت في تحذية أو تنسيق Alignment النص المكتوب بحسب قواعد معينة عادة تكون لمنع الردود المسيئة أو الضارة.(
- نافذة السياق :Context window رقم ثابت من الأجزاء اللغوية يمر على النموذج أثناء تعلمه فيفهم منه سياقا محددا .

2.2. التوطئة التاريخية

واعلم أن الكلام التاريخي مفيدٌ وممتع لأنه يساعد القارئ على تصور امتداد الفكر البحثي على مدى هذه العقود وطُرق تغيُّرِه وتطوره وعليه فسنشير إلى أهم النقاط الأساسية في تاريخ "معالجة اللغات الطبيعية" مشتملة على ما كان للعربية وهي على 5 مراحل أساسية [61][62][ع8] :

- المرحلة الأولى كانت فيما بين الخمسينيات والستينيات (1950-1960) وكانت أنظمة معالجتها مبنية تماما على القواعد المُهَندسة يدويا فتصاغ قواعد لغوية معينة ولها قسمان أساسيان المحللات النحوية أو المسترجعات القائمة على القواميس فمن أمثلة الأول نظام "إلزَا "Eliza للمحادثة وكان مبنيا على تحليل لغوي سطحي باستعمال "الأنماط الطبيعية "Regular Expression ومن أمثلة الثاني التصنيف المشاعري القائم على كلمات محددة وفي اللغة العربية (وكانت متأخرة قليلًا) بُنيت بعض الأنظمة الصرفية والاشتقاقية، ويحسن التنبيه إلى أن في هذه الفترة ظهرت أول أبحاث هيكلية النماذج العصبية بـ"المدرك "Perceptron على يدِ "رُزِبلَت ."Rosenblatt

- المرحلة الثانية من الثمانينيات إلى أواخر التسعينات (1980-1990) وفيها أطلت أوائل الطرق الإحصائية مع نموذج مَركوف الخفي (المستور) HMM وبدأت معها تطبيقات أعقد كوسم أقسام الكلام والتعرف على الكيانات المُسماة وأما بالنسبة للعربية فانطلقت شركة صخر [ع7] وأسهمت في إنتاج تطبيقات لغوية كثيرة وكذلك بعض الدراسات الإحصائية على المعاجم العربية مثل الصحاح والتاج وبناء المُشَكِّلات والمترجمات وكذا دراسات إحصائية على القرآن.

- المرحلة الثالثة من التسعينات إلى بداية الألفية الثانية (1990-2000) وهذه قريبة من المرحلة الأولى في القواعد الجاهز ولكنَّ عُمقَ التطبيق اختلف فهنا رُكز على بناء معاجم متخصصة استُخدمت في تصنيف متخصص للمدونات النصية وفي العربية متصلا مع المرحلة الثانية بُني معجم تركيبي للتوليد والتحليل الصرفي والنحوي ويُشار أيضا أن في أواخر هذه المرحلة وبداية التالية ازداد الاهتمام بالعربية ومعالجتها بسبب الأحداث السياسية (أحداث أيلول -سبتمبر-) وصدر كذلك "بنك العربية الشجري PATB "ونمت حاجة جمع المدونات العربية الكبيرة .

- المرحلة الرابعة بين (2000-2017) وهي للمعالجة التضمينية بأن صارت النصوص والكلمات تُمثَّل على هيئة متجه رقمي (بنوعيه إما عدِّيٍّ أو كثيف) كثير الأبعاد تُستخدم في مهمات لغوية كثيرة مثل التصنيف والاسترجاع وغيرها ومن أشهر أساليبها TF-IDF وكذا تضمينات word2vec وهذا الأخير ومشتقاته كان له صدى واسع في الأبحاث العربية واسْتُعمل كثيرا في كثير من الدراسات والتطبيقات] 59,60.[

- وهذه المرحلة الخامسة والتي نحن فيها (2018- إلى الآن) بزغ نجم الهياكل العصبية بشكل عام وبهياكلها المميزة بشكل خاص (كنماذج CNN للمعالجة الصورية) وفي المعالجة اللغوية ظهرت نماذج RNN وأدَّت أداءً ممتازًا (ولو كان محدودا مُكلفا) في التوليد والفهم السياقي ثم في حوالي 2019 ظهر نموذج BERT ذا الهيكلية المرمزة واكتسحَ في مهمات الفهم كالتصنيف وقريبًا من تلك الفترة ظهر نموذج GPT-2 الذي كان نواةَ ظهور التطبيق الأشهر ChatGPT في عام 2022 ولن نذكر الجهود العربية هنا لأنَّا أفردنا لحصر واستقراء النماذج العربية فصلًا كاملا، ويجدر هنا أن نكتب ولو بشكل مختصر السببَ الرئيسي لاجتياح نماذج المحولات مجال المعالجة اللغوية خاصة ومنه إلى كل المجالات عموما .

أشرنا في المرحلة الأولى إلى صدور أول بحث عن العصبونات العصبية الحوسبية باسم "المُدرك" ولسياقات وأسباب تاريخية بحثية وحوسبية (يطول الكلام جدا بذكرها) ما بدأت تطبيقات النماذج العصبية تظهر بفعالية إلا عام 2012 بالانتقال من الخصائص المُستخرجة يدويا إلى تلك المُتَعلَّمة آليا في الشبكات العصبية والتي تطورت وصارت ذات طبقات عديدة ولكن أقصرَ عجزها عن فهم السياقات وتتابعها عن استخدامها في المعالجة الآلية توليدا وفهما وإن كانت هناك محاولات لذلك، ثم ظهرت هيكلية الشبكة العصبية التكرارية RNN فحُلَّت فيها مشكلة تذكر السياقات اللغوية المبنيِّ عليها فهم النصوص المكتوبة باستعمال الحالات الخفية وتحسنت معها أداء التطبيقات اللغوية وخاصة بتطويرها بهيكلية "الذاكرة القصيرة المدى المُطولة LSTM "ولكن بقيت مشاكل طول السياقات وأيضا فإنها بطيئة التدريب وعليه فقد احتيج إلى هيكلية تجمع سرعة التدريب (التوازي بشكل أساسي) مع فهم السياق فظهر مبدأ "الانتباه" والذي يسمح للأجزاء اللغوية بأن تتواصل فيحصل كلٌّ منها مما حولها على ما تريد إعطاء وأخذا فبُنيَت عليه نماذج المحولات واستعملت مع الشبكات ذات التغذية الأمامية وزامن ذلك تطور كبير في القوة الحوسبية فجرفت هذه الهيكلية كل القدماء وما زال تحسينها وتطويرها من عام 2017 إلى وقتنا مستمرًا .

3. الباب الثاني ﴿ــــي تقارير أداء النماذج اللغوية

في هذا الباب سنستعرض أهم طرق تقييم النماذج وتقسيماتها الأساسية ثم نتطرق إلى بحوث تقييم النماذج العربية.
"ما لا يقاس لا يُطَوَّر [4]"هي مقولة مهمة جدا في المجال العلمي التطبيقي إذ هو معتمد على نتائج قياس التجارب وما نحن بصدده (مجال النماذج اللغوية) لا بُدَّ فيه من قياس رقمي يُلحظ فيه (ولو بشكلٍ جزئي) أداء النموذج على المهام المتنوعة لفهم نقاط قوة وضعف النموذج لتحسين تواصله ولضمان أدائه.
واعلم أن تقييم أداء النماذج اللغوية يرجع لثلاثة أصول تم استخلاصها من دراسة مسحية عن تقييم النماذج [41] تُسرد ثم يُذكر أثناءها ما أضيف للعربية مما تختص به وأول هذه الأصول متعلق بما الذي يُقاس وثانيها بماذا (بِمَ) يُقاس وثالثها كيفَ يُقاس وهذا تفصيلها :

- الأصل الأول متعلق بالأمر المراد قياسه وهو على ست أقسام أساسية وهي مشتركة بين العربية وغيرها من اللغات من حيثُ المبدأ ولكنَّ المدونة المستخدمة لا بُدَّ تخلتلف بحسب الثقافة والمعرفة المختلفة .
  - قسم المعالجة اللغوية الطبيعية (الأداء اللغوي) وفيه أولا الفهم اللغوي ويشمل التصنيف والإدراك وثانيا التحليل المنطقي (الاستنتاج) وثالثا التوليد اللغوي ويشمل التلخيص والمحادثة والترجمة وإجابة الأسئلة ورابعا وأخيرا مبدأ الواقعية ويشمل مشاكل الهلوسة والكذب.
  - قسم الأخلاقيات: ويشمل الاعتمادية (بأن يُقاس أداء النموذج على مدخلات خارجه توزيعه الإحصائي المُتَعَلَّم) وثانيا الأخلاق [5]وتشمل مسائل التحيز العرقي والثقافي والسياسي وثالثا وأخيرا الموثوقية وهي قريبة من مبدأ الواقعية .
  - العلوم وتشمل الطبيعية (من حساب وهندسة وفيزياء وكيمياء وغيرها) والإنسانية (وتشمل التاريخ والشرعيات وغيرهما) وهي مُضَمنة في القسم الأول ولكنها تُفرد هنا لكثرة أنواعها
  - العلوم الطبية والقانونية: وهذه مفردة من القسم الثالث أيضا لحساسية موضوعها وعُسره .
  - قسم التطبيقات الفرعية: كمسائل التعليم وأنظمة الاقتراح وطُرق التخطيط وتوليد النصوص المهيكلة

ويجدر الإشارة إلى صدور عدد من الأوراق قاست (قَيَّمت) أداء ChatGPT خاصة على عدد من المهمات العربية ومن ذلك [43] فقيمته وقارنته مع Bloom على مهام الفهم والتوليد وكذا ورقة "تقييم "Taqyim قاست أداءه على وسم أقسام الكلام وتصنيف المشاعر والتلخيص والترجمة الصواتية [6]وغيرها] 44.[

[4]قائلها ويليام طومسون، بارون كلفن الأول (26 يونيو 1824 - 17 ديسمبر 1907)، عالم رياضيات وفيزيائي رياضي ومهندس بريطاني

[5]وهذا استعمال مجازي يُراد منه أن ينحوَ النموذج سلوكًا حسب القواعد الأخلاقية لمطوّريه.

[6]هذه ترجمة مصطلحTransliteration

- الأصل الثاني متعلق بأنواع المدونات الاختبارية وهي إما عامة شاملة أو متخصصة في فرعٍ ما .

وقد صدر لتقييم النماذج العربية عددٌ جيد من المدونات الاختبارية نذكر منها العامة ثم المتخصصة. أما العامة فمن ذلك مدونة AraTrust (الثقة العربية) وفيها عدد من جوانب الموثوقية كالأخلاقيات والإهانة والأمان الرقمي والصحة النفسية والجسدية (الحيوية) [45]. ومن المدونات كذلك مدونة LAraBench (المعيار العربي) وهي مدونة حووِلَ فيها بناء قاعدة معيارية لتقييم النماذج اللغوية العربية ومن اختباراتهم المعنى السياقي وبناء الجمل واستخراج المعلومات والترجمة والتصنيف والاستيعاب وغير ذلك [46]. ومن المدونات كذلك "الغفى "Alghafa فيها أسئلة اختيار من متعدد وقد درَّب أصحابها نموذجا حجمه 14 مليار معامل واختبروه عليها] 47 .[

ومن المدونات مدونة "أُركة "ORCA وفيها اختبار لأمور تُعتبر صعبة عند النماذج كالتشابهات المعنوية والتوقع الهيكلي [48] وهذه ومدونة "إدراك النماذج العربية "ANLU (فيها أيضا ما يتعلق بالدلالات المعجمية والاستدلال والمعرفة) [49] كلاهما مختص بجانب "فهم اللغات الطبيعية ."NLU وأما في جانب اختبار "توليد اللغات الطبيعية "NLG فهناك مدونة "دُلفين "Dolphinوفيها اختبارات للتشكيل وتحويل اللهجة وأجوبة المحادثات وتوليد العناوين والمحادثات وتصحيح الأخطاء النحوية [50]. وآخر ما صدر من المدونات العامة مدونة "بلسم "BALSAM وهي مدونة ضخمة شارك فيها أكثر من عشر جهات أكاديمية تختبر 57 مهمة مختلفة] 53.[

وهناك مدونات عامة التوجه مخصوصة النوعية (وهي الأسئلة) فمن ذلك مدونة "عقاد "AQAD وقد حوت أكثر من 17 آلاف سؤال متنوع [51] وكذلك مدونة "أسئلة عربية "ArabicaQA وهذه مدونة أسئلة ضخمة فيها أكثر من 89 ألفَ سؤال عام وأيضا أضافوا 3.7 ألفَ سؤال لا جواب لها [52]. وأخيرا هناك مدونة "مملى "ArabicMMLU[7] وهذه فيها حوالي 15 ألف سؤال في مواضيع علمية كثيرة (على صيغة خيار من متعدد) من مستويات تعليمية مختلفة من الابتدائية إلى الجامعية لقياس الفهم] 54.[

وأما المدونات المتخصصة فهي قليلة نسبيًّا، فمنها مدونة "الرياضيات العربية "ArMATH ففيها عدد من الأسئلة الرياضية المتنوعة [55] وكذا صدر مؤخرا اختبار المعرفة القانونية في النماذج اللغوية باسم "قياس القانونية العربية "ArabLegalEval وهي مدونة جُمعت من مصادر متنوعة] 56.[

ونشير أخيرا إلى أمور مستطرفة أحدها في مسألة القياس الثقافي إذ أن النماذج المشهورة تدربت غالبا على الإنجليزية وتطرَّقت للعربية لِما فأدى ذلك إلى أجوبة بكلام عربي وثقافة غربية وهناك ورقة لطيفة العنوان "أتشرب خمرًا بعد الصلاة؟" أشارت لذلك بأمثلة كثيرة [57] والأخرى أن هناك مدونات اختبارٍ عديدة اللغات منها العربية منها XTREME وهي منشورة [58] وأخيرا هناك قائمة أوائل لقياس أداء النماذج على عدد من المدونات السابقة على موقع "ضَمَّة "Hugging Face يُمكن رفع وتجربة النماذج عليها .[8]

- الأصل الثالث وهو مختص بالكيفية أي "كيفَ يُقاس" وهو نوعان إما آلي (مؤتمت) أو إنساني والأول محدد

[7] تسمية لطيفة وكأنه مكان إملاء على وزن مَفعَل والإملاء من الاختبارات المشهورة.

[8] وهذا رابطها https://huggingface.co/blog/leaderboard-arabic :

أما النوع الآلي (أو قُل الرقمي) فله عدة مقاييس تُستخدم مع عدد كبير من المدونات ونلخص هنا أشهرها [42]:

(1) مقاييس تتعلق بالدقة التصنيفية (الفئوية) تُحدد فيها أولا النتائج على أقسامها الأربعة[9] ثم تحسب الدقة "تقيس نسبة التوقعات الصحيحة من بين جميع التوقعات".

$$Acc = \frac{TP + TN}{TP + TN + FP + FN}$$

ويحسب كذلك "الاستدعاء" (أو قل المعدل الإيجابي الصحيح) وهو نسبة الإيجابيات الفعلية التي حُددت تحديدا صحيحا

$$Recall = \frac{TP}{TP + FN}$$

ويُحسب "الإحكام" وهو نسبة الإيجابيات المحددة التي كانت صحيحة فعليا

$$Precision = \frac{TP}{TP + FP}$$

وبعد ذلك تُحسب نسبة مستخدمة كثيرا (وستأتي في وصف النماذج) وهي مقياس فاء 1 F-1 score-وهو مقياس لدقة النموذج في التصنيف الثنائي عن طريق حساب المتوسط التوافقي للإحكام والاستدعاء وله نوع آخر متعلق بالوسيط (الوسط) التوافقي للإحكام والاستدعاء.

$$F1 = \frac{2 \times Recall \times Precision}{Recall + Precision}$$

وهذه أهم المقاييس التصنيفية ويراجع غيرها في المصادر المطولة [42].

(2) مقاييس تتعلق بتشابه الأجزاء اللغوية وهذه المقاييس عادة تستخدم في مهمات التوليد النصية كالترجمة والتلخيص نظرا إلى أن المخرج له مقابل حقيقي (صحيح) وأشهرها التالي:

مقياس BLEU وهو مقياس يقيس التشابه بين النص المُولّد والنص المرجعي باستخدام أجزاء لغوية محدودة العدد والقيم الأعلى تشير إلى أداء أفضل.

$$\text{BLEU} = \text{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right)$$

حيث أن BP تشير لعقوبة طول المُخرج p_n هو دقة الجزء اللغوي و w_n يشير للوزن الاحتمالي للجزء اللغوي .

والثاني مقياس :ROUGE يقيس التداخل بين بين الأجزاء اللغوية بين النص المُولّد والنص المرجعي :

$$\text{ROUGE-N} = \frac{\sum_{\text{ref}} \sum_{\text{ngrams}} \min(\text{count}_{\text{gen}}, \text{count}_{\text{ref}})}{\sum_{\text{ref}} \sum_{\text{ngrams}} \text{count}_{\text{ref}}}$$

---

[9] أذكرها في الحاشية حتى لا يطول السياق وهي: مُخطئ إيجابا (FP) وسلبا (FN) وهذه إن توقع شيئا خطأ وهنا مُصيب إيجابا (TP)وسلبا (TN) أو بتعبير آخر: إيجابي وسلبي خاطئ وسلبي وإيجابي صحيح .

ولها أيضًا أنواع أخرى من الأنواع مثل مقياس "الارتباك PPL "(أو قل التشويش) وهو متعلق بمدى تأكد النموذج من احتمالية مُخرَجِه بناءً على مخرج مرجعي (حقيقي) يختاره المطور (الباحث) وهذه معادلته ولا شك أن المراد أن تقل هذه النسبة.

$$\mathrm{PPL} = 2^{-\frac{1}{N}\sum_{i=1}^{N}\log_2 P(x_i)}$$

والاحتمال P لـ x_i (وهي جزء لغوي) أي احتماله المتوقع من النموذج.
وهنا مسألة ينبغي الانتباه لها أن اختبار المُخرج إما أن يكون احتماليا بأن تُحسب احتمالات المخرج ومقارنتها مع المراد توليده فيظهر الفرق قُربًا أو بُعدًا ومن أشهر ذلك مدونة MMLU أو بالمقارنة النصية كما سبق وأشهر أمثلته الترجمة والتلخيص .

والنوع الثاني من طرق التقييم هو التقويم البشري وهو نسبي (وقد استخدم كثيرا في النماذج العربية نظرًا لقلة مدونات الاختبار كقياس الأسلوب الكتابي وهو أمر نسبي وهناك أمور واضحة كالأخطاء العرفية أو اللغوية) وسيأتي ذكرُه كثيرا عند سرد النماذج العربية .

4. الباب الثالث في إحصاء النماذج العربية الضخمة

1.4. تمهيد

في هذا الباب سنتكلم عن النماذج الكبيرة والتي طُورت بهيكلية المفكك فقمنا بفصل الكلام عليها وتفصيله واختصار الأوراق البحثية التي كتبها ناشروا النماذج لأنها النماذج الأكبر وهي التي يقع عليها مصطلح النماذج الضخمة وذكرنا معها أيضا النماذج الصغيرة إما لفائدة تاريخية أو بحثية والكلام في هذا المفصل منقسم إلى قسمين الأول قسم مختصر وفيه جدول يحتوي وصفًا مختصرا لكل نموذج وتاريخ إصداره وبعض المعلومات الأساسية عنه ثم القسم الثاني وفيه تفصيل عن كل نموذج مختصر من الورقة البحثية أو التقرير المنشور عنه. ثم نتبع المفككات بالكلام عن نوعَي الهياكل الباقية وهي المرمزة و المرمزة المفككة ونعرض فيها أشهر النماذج العربية ونشير لغالبها.

2.4. نماذج المحولات المفككة (القسم المُختصر)

اسم عائلة النموذج	تاريخها	المصدر	عدد النماذج	وصفٌ مختصر
عائلةAraGPT	2020	الجامعة الأمريكية في بيروت	٤ نماذج	أول عائلة نماذج عربية تظهر وهي منشورة على الشبكة [10]أُعتمدت فيه هيكلية المفكك (فاك التشفير). دُربت على 77 جيجابايت من النصوص المجموعة (حوالي 9 مليارات كلمة) غالبها المقالات، واستعمل فيها مقسم.BPE
عائلةBloom	2022	شركة بيج ساينس BigScienceبالتعاون مع جهات أخرى.	نموذج واحد أساسي (وخمسة ثانوية)	نموذج عديد اللغات منها العربية منشور على الشبكة [11]أهم نسخه حجمها 176 مليار معامل (وله نسخ أخرى أصغر) دُرّب على 46 لغة بشرية و 13 لغة برمجية. أُستعملَت فيه هيكلية المفكك، وبلغ حجم مدوَّنته النصية كاملةً 1.6 تيرابايت نسبةُ العربية منها 4.6% ما يقارب 75 جيجابايت.
عائلة ياسمين Jasmine	2022	جامعة بريتش كولومبيا British Columbia وجامعة محمد بن زايد MBZUAI	أربع نماذج	عائلة نماذج عربية تتراوح أحجامها من 300 مليون إلى 6.8 مليار عامل دُربت على 235 جيجابايت من النصوص العربية الفصيحة فقاربت 25 مليار جزء لغوي غالبها نصوص من الشبكة. استعمل فيها مقسم .BPE ونُشر فقد أصغر نسخها.[12]
عائلة AceGPT الأولى والثانية	-2023 2024	معهد شنجن Shenzhen الصيني للبيانات الضخمة و جامعة هونج كونج في جنشن Shenzhenوجامعة كاوست في السعودية	نموذجان للأول واثنان للثاني	عائلة AceGPT لها نسختان كل واحدة منها في نموذجان بسبعة وثلاث عشرة مليار (7م + 13م) كلها إعادة تدريب لنموذج لاما-2 لتعليمه العربية (هو نموذج انجليزي) فدُربت على حوالي 30م جزء لغوي ثلثاه تقريبا عربي واستُعملَ في تدريبه أيضا الضبط الدقيق والتعلم المعزز وكلها منشورة على الشبكة[13]

جدول 1: جدول النماذج المفككة

[10]الرابط: https://huggingface.co/aubmindlab

[11]الرابط: https://huggingface.co/bigscience

[12]الرابط: https://huggingface.co/UBC-NLP/Jasmine-350M

عائلة جيس Jais الأولى والثانية	-2023 2024	شركة انسبشن Inceptionوجامعة محمد بن زايد وشركة أنظمة كرباس Cerebras Systems	نموذجان للعائلة الأولى وأكثر 20 للعائلة الثانية	هذه العائلة (13 مليار و30 مليار للأولى وأحجام وهياكل متنوعة للثانية تتراوح بين 300 مليون إلى 70 مليار) تعتبر أحد أكبر النقاط البحثية للنماذج العربية لاشتمالها على فتوحات انفردت بها فيما يتعلق بحجم البيانات المتدرب عليها (بلغت حوالي 72 مليار جزء لغوي عربي يشمل المترجمات وكذا تحتوي الإنجليزية) وطرقها والعتاد الحوسبي وكذلك ضُبطَ ضبْطًا دقيقا وكل هذه النماذج منشورة على الشبكة[10]
نموذج نونNoon	2024	شركة نون في السعودية	نموذج واحد	نموذج دربته شركة نسيج مبني على نموذج Bloomله 7 مليار معامل وهو المنشور على الشبكة [11]وليس له تقرير منشور.
عائلة آيةAya-23	2024	شركة كوهير الأمريكية Cohere	نموذجان	هذه آخر مُخرجات سلسلة النماذج عديدة اللغات التي بدأتها شركة كوهير وهما نموذجنا بحجم 8 مليارات و35 مليار متغير لثلاث وعشرين لغة (منها العربية) وكلها مفتوحة المصدر [12]ولها إطار استعمال برمجي API[13] حصدت نتائج مرتفعة في تقييماتها.

---

[13]الرابط https://huggingface.co/FreedomIntelligence :

[10]على الرابط https://huggingface.co/inceptionai :

[11]الرابط https://huggingface.co/Naseej/noon-7b :

[12]الرابط https://huggingface.co/CohereForAI :

[13]الرابط https://docs.cohere.com/docs/chat-api :

عائلةArabianGPT	2024	معمل الروبوتات وإنترنت الأشياء RIOTU في جامعة الأمير سلطان	أربعة نماذج أساسية	أربعة نماذج صدرت من جامعة الأمير سلطان بأحجام مختلفة: 134 مليون و 345 مليون و800 مليون و1.5 مليار معامل. وكل منها صدر منه بعض النسخ المضبوطة، وكل ذلك مفتوح المصدر [14]وكذلك أُصدرت مكتبة للمقسمات المدربة وهي متنوعة ما بين بتشكيل وبدونه.[15]
نموذج علامALLAM	2024	الهيئة السعودية للبيانات والذكاء الاصطناعي (سدايا SDAIA)والمركز الوطني للذكاء الاصطناعي (SCAI)	نموذج واحد	أحدث ما صدر من النماذج العربية (ويشمل الإنجليزية) وله نسختان نسخة أصلها لاما-2 وضُبط بعدها على العربية وحُسِّنَ مقسمه (لها ثلاث أحجام 7م, 13م, 70م) ونسخة دُرِّبَت من الصفر حجمها 7م. ضُبط ونُسِّقَ (alignmented).
عائلةQWEN2	2024	شركة علي بابا وفريق كوين Qwen Team, Alibaba Group	خمسة نماذج	من النماذج الإنجليزية التي احتوت على عددٍ آخر من اللغات منها العربية أحببت الإشارة له لأنه مفتوح المصدر واستعملته ورقة علام في بعض التقييمات وهو مفتوح المصدر[17] [16]
نموذج سلمىSilma	2024	شركة سلمى المغربية	نموذج واحد	هذا أحدث نموذج عربي صدر إلى الآن وهو المتصدر على قائمة النماذج العربية [17]وأصله نموذج جيما Gemma من جوجل ولا تتوفر معلومات كثيرة عنها ولكن من الملفات المنشورة يبين أنه نموذج جيما-2 نفسه ولكن تم تدريبه على نصوص عربية[18]

[14]الرابط: https://huggingface.co/riotu-lab

[15]الرابط: https://github.com/riotu-lab/aranizer

[16]الرابط https://huggingface.co/Qwen: وأثناء كتابة هذه الأوراق صدرت نسخة جديدة 2.5 منه وهي على نفس الرابط.

[17]الرابط: https://huggingface.co/spaces/OALL/Open-Arabic-LLM-Leaderboard

[18]الرابط: https://huggingface.co/silma-ai/SILMA-9B-Instruct-v1.0

3.4. نماذج المحولات المفككة (القسم المُطَوَّل)

في هذا القسم سنفصل عن كلِّ نموذج تفصيلا يتضمن غالبَ فوائد ورقته البحثية (إن وُجِدَت) مما يُفيد شادي التدريب والتطوير والبحث مما يُستعمل في تدريب النماذج وكذا طُرُق جمع المدونات ومعالجتها فيُوَفِّرُ عليه عناء قراءة عشرات الصفحات، وهي مرتبة هنا على نفس ترتيب الجدول السابق

- عائلةAraGPT

نماذج عائلة AraGPT2 تأتي بأربعة أحجام مختلفة مع اختلافات في المحسّنات وبُعد التضمين وعدد الرؤوس والطبقات .
فنموذج AraGPT2-base بحجم 135 مليون معامل، يحتوي على نافذة سياق بطول 1024 وحدة لغوية وبُعد تضمين 768 لكل وحدة، ويضم 12 من رؤوس الانتباه و12 طبقة، ويستخدم مُحسّن.LAMB
ونموذج AraGPT2-medium بحجم 370 مليون معامل، يتضمن نافذة سياق بنفس الطول ولكن ببُعد تضمين 1024 وعدد رؤوس الانتباه 16 وعدد الطبقات 24 .
أما نموذج AraGPT2-large بحجم 792 مليون، فيحتوي على نافذة سياق 1024 وبُعد تضمين 1280 ورؤوس الانتباه 20 وعدد الطبقات 36، ويستخدم مُحسّن .Adafactor
أخيرًا، نموذج AraGPT2-mega بحجم 1.46 مليار، يحتوي على نافذة سياق 1024 وبُعد تضمين 1536 ورؤوس الانتباه 25 وعدد الطبقات 48.
أُستخدم في تدريب النماذج (على وحدة معالجة تنسر واحدة (TPU مدونة نصيةً حجمها 77 جيجابايت فيها 8.8 مليار كلمة، (جُلُّها منشور) غالبها مقالات عربية فصيحة وفيها أيضا نصوص مسحوبة من الشبكة و ويكيبيديا (الموسوعة الحرَّ)، وعولجت معالجة مبدئية قبل أن تُستخدم في التدريب، واستخدم في تقسيمها المقسم BPE وبلغ عدد وحداتها اللغوية المميزة 64000 .
دُرِّبَ النموذج الأساسي والأوسط (Base + Medium) باستخدام مكتبة gpt-2-simple[19] وأما النموذج الكبير والعملاق (Large + Mega)باستخدام مكتبة .gpt2-ml[20]
قُيِّم النموذج بأسلوبين الأول باستخدام أسئلة منوعة أجاب عن ربعها بشكل صحيح (25%) وفشل في مهمة الترجمة نظرا لقلة (شبه انعدام) النصوص الإنجليزية من المدونة التدريبية، والأسلوب الثاني فهو بالتقييم البشري حيث عُرضت مجموعة من النصوص المولدة من النموذج مع عدد من النصوص الطبيعية (ثمان من هذه وثمان من تلك) على 74 شخص عربي وبحسب الورقة فإن المُولّدات خدعت 60% من عينة الاختبار ويشار أيضا إلى أن 50% منهم صنفوا النصوص الطبيعية كمولدة] .1 .[

وهذه العائلة أول تجربة في تدريب النماذج العربية باستعمال المحولات المفككة وكانت نواة وبذرة الانطلاق.

[19] وهذا رابطهاhttps://github.com/minimaxir/gpt-2-simple

[20] وهذا رابطها/https://github.com/imcaspar/gpt2-ml

- عائلة بلومBloom

نموذج بلوم Bloom هو نموذج عملاق يبلغ حجمه 176 مليار معامل طورته شركة بيج ساينس وعمل عليه مئات الباحثين .[21] دُرّبَ هذا النموذج الضخم على مدونة نصية مهولة بلغ حجمها 1.6 تيرابايت تحوي 46 لغة بشرية متنوعة جدا كالإنجليزية والصينية واللغات الأوروبية والعربية وبعض اللغات الإفريقية وغير ذلك ونسبة العربية من ذلك 4.6% (تبلغ تقريبا 75 جيجابايت) ،[22]جُمعت هذه المدونة بمجهود فريق المدونة بطرق كثيرة منها ما كان في مسابقات برمجية متنوعة وكذا بالبحث على الشبكة على المدونات المجموعة فللعربية استعملوا مدونة "مصادر" والتي قامت بجمع معلومات وروابط عدد كبير من المدونات العربية.[23]

وكذا استعملت نصوص مجموعة من الشبكة (من مدونة اسمها (OSCAR عولجت بمنهجيةٍ خلاصَتُها تدور رحاها حول حذف المستندات المتكررة وذات الترميزات غير المفيدة والكلمات البذيئة وذات نسبة الارتباك (التشويش) العالية، فبلغت نسبة العربية المحذوفة بعد التصفية 20% (فليُعلم أن هناك ورقة بحثية كاملة عن المدونة النصية [3.([

اِختيرت هيكلية المفكك لهذا النموذج بسبعين طبقة (70) كل منها فيه 112 رأس انتباه وبُعد تضمين جزئه اللغوي 14336 ونافذة سياقه 2048، والتضمينات الموضعية استُعملت لها خوارزمية .ALIBI أما المقسم فقد أُستعمل المقسم BPE بأجزاء لغوية مميزة بلغت 250,680 وقد وضعوا فيها مئتي جزءٍ إضافي للاستخدامات المستقبلية.

لتقييم النموذج صُمِّمَت منهجية متكاملة تشمل الأسئلة والترجمة والتلخيص وكذا توليد النصوص البرمجية واختبار جودة ما ينتجه النموذج من تضمينات المستعملة في عدد من التطبيقات. ففي الترجمة من العربية للإنجليزية بلغت نسبة الصحة (على معيار بلو (BLEU حوالي 40%. وأما في التلخيص فبلغت صحته (على معيار (ROUGE-2 ما يزيد عن 10 ،[24]وأما في التضمينات فبلغ دقة العربية في مهمة التصنيف 59.25 وفي مهمة التشابه المعنوي 58.67.[25]

ويحسن ختم هذه الفقرة بالإشارة إلى أن هذا المشروع مفيدٌ جدا الاستفادة منه في بناء النماذج العربية العملاقة إذ في تقريرها [2] معلومات مفيدة جدا عن التدريب يشابهه ما أصدرته ميتا عن نماذجها الجديدة لاما 3.1 [4] إذ تفاصيل التدريب التقنية تكاد تكون أهمّ ما يتعلق بهذه المشاريع.

---

[21]ومن اللطائف أن أول صفحتين ونصف من ورقة النموذج كلها في أسماء من شارك إما في المدونة أو المقسم أو هندسة الأوامر أو الهيكلية أو التقييم وغيرها من الأقسام، وهذا يَدلُّ على ضخامة هذه المشاريع وحاجتها للموارد البشرية الكبيرة.

[22]وهذا الحجم صغير مقارنة بالنماذج الجديدة ولكن هذا النموذج من أوائل ما صدر في هذا المجال ضخمَ الحجم عديدَ اللغات.

[23]رابطها: https://huggingface.co/datasets/arbml/masader

[24]المعذرة عن عدم ذكر الرقم الدقيق إذ لم يُذكر في الورقة سوى رسم بياني .

[25]في الورقة استعملوا هنا نسخة السبعة مليار من النماذج ولم يذكروا نتائج النموذج الأساسي.

- عائلة ياسمين Jasmine

عائلة نماذج ياسمين Jasmine مكونة من أربعة نماذج تتبع هيكلية المفكك أحجامها كالتالي:
أصغرها بـ350 مليون معامل له 12 طبقة و12 رأس انتباه وبُعد مضمنها 768، والثاني 1.3 مليار معامل له 24 طبقة و16 رأس انتباه وبُعد مضمنها 2048،والثالث 2.7 مليار عامل له 32 طبقة وكذا رأس انتباه وبُعد تضمين 2560، وأكبرها 6.7 مليار معامل كنفس السابق إلا إن بُعد تضميناته 4096.
طور هذه النماذج مجموعة التعلم العميق ومعالجة اللغة الطبيعية في جامعة كولومبيا البريطانية British Columbia وكذا قسم معالجة اللغة الطبيعية والتعلم الآلي في جامعة محمد بن زايد MBZUAI مستعملين في ذلك مكتبة.[26]GPT-Neo
دُرِّبت هذه النماذج على مدونة نصية كبيرة مقارنة بالعائلتين السابقتين إذ بلغت في مرحلة التدريب الأولي حوالي 244 جيجابايت (23.7 مليار جزء فصيحٍ و13 عامي من تويتر) غالبها كان منخولا من نصوص الشبكة وتنوع الباقي بين مقالات اخبارية وكتب ومقالات ويكيبيديا .
قُيّمت هذه النماذج في خمس مهمات أساسية :نمذجة اللغة، والإكمال التلقائي والإدراك العام والتلاعب بالكلمات وفهم اللغة الطبيعية. أما في نمذجة اللغة فقد حسب الباحثون ارتباك النماذج لغويا فحصد أكبرها 42.25 في متوسط أداءه على عينات نصية جُمعت من ويكيبيديا وبعض المقالات الإخبارية والكتب. وأما في اختبار الإكمال التلقائي فقد اختبرت النماذج على عينة من 15 ألف مثال ثلثها في عناوين المقالات الإخبارية وثلثها في عناوين أطروحات علمية وثلثها في قصص إخبارية ثم طلب من النماذج إكمال النص بعد إعطائها جزءًا من العينة وحُسب أدائهم بمقياس فاء-1 وكذلك تقدم أكبر النماذج الياسمينية بأعلى نتيجة.
وفي تقييم الإدراك العام فقد قام الباحثون بعمل مدونة مخصوصة لهذه المهمة نظرا لانعدام واحدة عربية بمنهجية ذات ثلاث مراحل أولا جمع عينات من موقع WikiHow ومن ثم ضبط نموذج AraT5 على عينة من عشوائية من العينات السابقة وولدوا لها أجوبة ثلاثة لكل سؤال فصاروا أربعة مع الصحيحة الأصلية وثالث المراحل لزيادة صعوبة الاختبار قاموا باستعمال أسلوب "التصفية العدائية" وضبطوا نموذج MARBERT ليصنف ما سبق من الأجوبة ما بين حقيقي ومولد ليُنتجوا مدونة من الأسئلة قُيِّمَت عليها النماذج الثلاث الأُوَل بمقياس الدقة وكذا الفاء-1 (بما يُسمى بنتيجة النموذج اللغوي (LMS فحصد ذي 2.7 مليار معامل [27]أعلى النتائج بـ37 في المقياسين.
وفي تقييم التلاعب بالكلمات فاختُبِرت النماذج على خمس طرق تلاعب يراد إرجاع الكلمات الأصلية منها أولها أن تدار أحرف الكلمة [28]والثانية بأن تغير حروف الكلمة عشوائيا سوى الحرف الأول والأخير والثالثة نفس الثانية ولكن ثُبِّتَ الحرفان بدلَ واحد والرابعة إضافة عشوائية لمسافة أو علامة ترقيم والخامسة بأن تُعكس حروف الكلمة واستعمل مقياس فاء-1 ونتائجه ضعيفة جدا. وأما تقييم فهم اللغة الطبيعية فقد قُيِّمت النماذج على مهام تصنيف ستة بتدرج عدد الأمثلة من واحد إلى 24 اكتسحها أكبر النماذج 6،7.
أخيرا فقد قُيم النموذج تقييما بشريا في توليد الأخبار على عينة مُقَيّمين ونتائج ذلك قاربت أن تكون اختيارا عشوائيا (احتماله 50% لكل خيار أموَلَّدٌ أو حقيقي)، وكذا قُيم في كتابة الشعر بعد ضبطه على مجموعة شعرية فيها 22 ألفَ بيت

[26]وهذا رابطها: https://github.com/EleutherAI/gpt-neo

[27]لم يختبروا أكبر النماذج في هذا التقييم.

[28]لم يبين كاتبو الورقة بالضبط ما المقصود بتدوير الكلمة ولكن حسب الفهم يظهر لنا أنها تغير أماكن كلماتها بعضها في بعض وذكروا مثالا "الجيولوجي" تصير "يولوجيالج" دونَ إضافة أحرف أو حذفِ أخرى.

وحسب الورقة فإن المقيمين يعرفون المولد بنسبة 52.63% وكذلك ضُبط النموذج على التغريد بسابقة "غرد:" قبل كُلِّ تغريدة واستطاع المقيمون تحديد المولد بنسبة 48.53% وكذلك قُيِّم أداء النموذج على توليد الكلام العامّي بخمسٍ من لهجاته وهن المصرية والجزائرية والأردنية والمغربية واليمنية وظهر أن أداء النموذج جيدا في الأولى فقط وضعيف في الباقيات. ختم الباحثون اختباراتهم بدراسة بعض جوانب التحيز الاجتماعي كوقوعه بين سياقات الجنسين وكذا اللون والديانة وأظهرت نتائجهم الأولية "أن هذه النماذج ولو تدربت على مدونات متنوعة فلا بد أن تعاني تحيزات متفاوتة] "5.[

وهذه الورقة فيها فوائد كثيرة لكثرة التجارب والاختبارات التي اسْتُعملت فيها.

- عائلة أَيس الأولى والثانيةAceGPT 1,2

عائلة AceGPT 1,2 من تطوير جهاتٍ صينية بالتعاون مع جامعة الملك عبد الله (كاوست) لها نسختان كل واحدة منها في نموذجين بسبعة وثلاث عشرة مليار (7م + 13م) كلها إعادة تدريب لنموذج لاما-2 ولنبدأ بالأولى:

كان هدف هذه النماذج هو توطين (أو قل تعريب) النماذج الكبير المدربة (وفي حالتنا لاما-2) على اللغة العربية بعد ملاحظة مشاكل إدخال النصوص المترجمة على المدونات التدريبية (إشارة إلى نموذج جيس وسيأتي).

استعمل في تدريب النموذج بيانات عربية نسخة 2022 [29] على التوزيع التالي (حرف الميم هنا اختصار للمليار) :

لاما-2-7م درب على 30م جزءٍ لغوي 19م منها عربية والباقي انجليزي وأما نسخة لاما-2-13م فدربت على 6م جزءٍ عربي و4م جزء انجليزي ولم تُستعمل كل المدونة نظرًا للمحدودية الحوسبية والتكلفة العالية .

من المهم الإشارة إلى أن نماذج لاما-2 لا تحتوي الا على الرموز العربية الأساسية المتضمنة الحروف وبعضًا آخر كلها تصل إلى 53 رمزًا ولم تُكَثَّر في النموذج الأول بل في الثاني (وسيأتي) .

بعد التدريب الأولي (وهو على البيانات السابقة) دُرِّبَ النموذج تدريبا ضبطيًّا (أي بالضبط الدقيق) استعملت فيه أسئلة عربية جمعت من موقع كورا Quora واختير نظرا لكثرة الأسئلة الدالة على ما يهتم به الشعب العربي وللإنجليزية استعملوا عددا من المدونات منها Alpaca وترجموا كذلك مدونة ShareGPT آليا وأعادوا توليد النتائج حرصا على عدم وجود مشاكل ثقافية .

ثم التدريب بالتعلم المعزز استعملت فيه تغذية راجعة من الذكاء الاصطناعي بدلا من البشر نظرا لغلاء كِلفتها واختيرت خوارزمية PPO، جمعوا لها 40 ألف سؤال متنوعة ثقافيا ودينيا وقيَّم المخرجات ChatGPT-4 بعد تهيئته بأوامر معينة .

قُيِّمت نتائج النماذج على أسئلة صواب وخطأ عددها ثمانية آلاف (8000) وُلِّدت من ChatGPT-3.5 وراجع جُمَيعة (مجموعة جزئية) منها بشر. أما النتائج مقارنة مع نماذجَ أُخَر فبلغ أداء AceGPT على المجموعة المراجعة من الأسئلة 74.62% بمقياس فاء-1 مقارنة بأداء GPT-3.5 البالغ 79.03% وكذلك قُيمت هذه النماذج على أسئلة عامة وأسئلة معرفية من مدونة MMLU و EXAMS ومدونة عربية جمعوها اسمها ALUE (مُقَيِّمة فهم اللغة العربية] 8[) وقاربت نتائج النموذج 13م لهم نتائج نموذج AraMUS(ويشار له في النماذج المرمزة المفككة) وما سبق خلاصة معلومات العائلة الأولى] 6.[

وأما العائلة الثانية فانطلقوا من مبدأ "اكتساب اللغة الثانية" وعلى نفس نماذج لاما-2 ولكن انتهجوا نهجا جديدا في زيادة عدد الأجزاء اللغوية ومع أنها "متواضعة الأداء" مستعملين خوارزمية IBPE[30] (وهي المعتادة) [9] حيثُ قاموا بالتدريب على مراحل متصاعدةٌ فيها عددُ الأجزاء اللغوية المضافة باطراد مع نسبة اللغة العربية المُتَدَرَّبِ عليها وبانعكاس مع نسبة الإنجليزية ومجموع بيانات التدريب الكلية 30م جزء لغوي .

أما في الضبط الدقيق فقد استعملوا مدونة جمعوها واسمها ALAN (الضبط الدقيق العربي للنماذج اللغوية) بطريقة لطيفة حيثُ حصروا 127 موضوعا عامًّا عربيا وشققوا (فرعوا) منه 11 ألف عنوانا ثم جَزَّؤُوها لـ244 ألف نقطة معرفية وولدوا بها باستعمال ChatGPT-3.5 ما بلغ 733.419 نص للضبط الدقيق .

أما بيانات التدريب فاستعملوا 2368 كرتَ شاشة [31]ومحسن AdamW ونافذة سياق بطول 4096 جزء لغوي وأما التقييم فغالب معلوماته تُشبه النموذج الأول ولن أطيل بها تراجع في تقرير النموذج] 7.[

---

[29]المنشورة على الشبكة باسم ArabicText 2022 وهي من جمع جهة صينية

[30]هي تمديد للأجزاء اللغوية الناتجة عن استعمال خوارزمية BPE بعدَ تدريبها لتوسيعها على لغة أخرى

[31]لم يبينوا نوعه

- عائلة جيس الأولى والثانيةJais 1,2

عائلة جيس الأولى صدر أولا فيها نموذج حجمه 13 مليار بنسخة عادية (مولِّدَة) ونسخةٍ مضبوطة (مدردشة (ChatBot وهو الذي تكلمتْ عنهُ ورقتهم المنشورة [10 .]

استخدم الباحثون مدونة نصية عربية بلغ حجمها 72 مليار جزء لغوي كان جزءٌ منها (18 مليار) مترجما من ويكيبيديا (3 مليار جزء) والكتب (15 مليار جزء) والباقي من المدونات المنشورة وغالبها مسحوب من الشبكة ثم قام الباحثون بزيادة العينات [32]لتصير 116 مليار جزء وأما الانجليزية فاستعملوا 232 مليار جزءٍ فيها 46 مليار نص برمجي من جتهب .Github عولجت النصوص المجموعة بازالة الشوائب الرمزية والتشكيل وحذف القصيرة جدا منها .

بيَّن الباحثون أنهم استخدموا حجمًا كبيرا (13 مليار) للنموذج نظرا إلى أن "النموذج الأكبر يفهم أكثر" واضطروا حسب قوانين تحجيم النماذج [33]أن يستخدموا الانجليزية لزيادة بيانات التدريب .

استعمل مقسم BPE وعدد أجزائه المميزة 84,992. في هيكلة النموذج استعملت متجهات ALiBi ودالة تنشيط SwiGLU ونافذة سياق بطول 2048 و 40 طبقة و40 رأسَ انتباه وبُعد التضمين 5120 ومحسن .AdamW وقد دُرِّبَ النموذج على حاسوب خارق اسمه ."Condor Galaxy"

ضُبط النموذج على 3 ملايين عينة عربية أكثرها مُترجم و6 ملاين أخرى إنجليزية وبُنيَت النصوص إما على صيغة السؤال والجواب أو المحادثة المطولة .

قُيّمَت النتائج أولا على معرفة الكلمات وثانيا على الإدراك العام وثالثا على التحيز والخطأ وفي تقييم الأجوبة العام استُعمل ChatGPT-4.

وفي مسألة الأمن أُنتبه لخمس أمور أساسية العنصرية وتسريب المعلومات والأذية والجريمة والتلاعب النفسي وهنا تنبيه أنهم أصدروا نسخة مثل السابقة بحجم 30 مليار قبل العمل على العائلة الثانية

والتي فيها فرعان الأول "jais-family" دُرِّب مثل الأول تتراوح أحجامه من 590 مليون إلى 30 مليار (نافذة سياقه بلغت 16000 جزء) دُربت مثل النموذج الأول ولكن على مدونة جديدة حجمها بلغ في العربية 490 مليار جزءٍ لغوي وأما الفرع الثاني فهو يشبه AceGPT إذ قاموا بتدريب نموذج لاما-2 "jais-adapted" على ما يقارب 320 مليار جزء لغوي أضيف للنموذج هنا 32000 ألف جزءٍ جديد عربي وقُيّمَت النتائج على مدونات مترجمة قالوا أنه قد تم مراجعتها من قبل مراجعين عرب] .11[

---

[32]لم يذكروا تفصيلَ ذلك ولكن على ما يبدو أنه تكرار لبعض النصوص

[33]ما هو معروف بscaling laws

- نموذج نون

نموذج نون من شركة نسيج السعودية. استخدموا فيه نموذج ،BLOOM 7B وهو نموذج بخصائص محددة حيث يحتوي على بُعد تضمينات يبلغ 4096، وعدد رؤوس الانتباه 32، ويضم 30 طبقة. وله نافذة سياق بطول 2048 وحدة لغوية، مع استخدام .ALiBi يحتوي مُقَسِّمُه على 250,880 جزء لغوي. بالرغم من أن الباحثين لم يذكروا تفاصيل البيانات المستخدمة أو منهجية التدريب، فإن هذه المعلومات تم استخراجها من ملفات النموذج الداخلية المنشورة على.Hugging Face

- نموذج آية-23

نماذج آية-23 هي آخر ما قدَّمته شركة كوهير في سلسلة نماذجها عديدة اللغات والتي بدأت بنموذج آية-101 والذي اعتمد هيكلية المرمز المفكك ونموذج mT5 كأساس تم ضبطه دقيقا ونُشر وله تقرير كامل [14] وبعدها نُشرت نماذج Command R والتي اعتمدت هيكلية المفكك وكان حجر أساس نماذج آية-23 إذ وُجدت مشاكل في آية-101 منها قِدَمُ هيكليته ومعرفته وضُعْفِ أداءه وشساعة لغاته وكثرتها مما أتعبه "بلعنة التعددية اللغوية" فارتُئِيَ تركيز الاهتمام والموازنة بين الاتساع والتعمق فحددت 23 لغةً ،

فصدر نموذجان جديدان آية-23-8م (مليار) وكذا آية-23-35م معتمدا على ضبط نموذج Command R واختير في هيكليته: مبدأ الانتباه المتوازي ودالة تفعيل SwiGLU وتضمينات RoPE ومقسم BPE بعدد أجزاء لغوية بلغ 256 ألفًا وفي نموذج 8م فاستخدمت مُستعلمات الانتباه المُجَمَّعة.Grouped Query Attention

ومواصفات النموذجين كالتالي: أما الصغير فبُعد التضمين 4096 وطبقاتها 32 ورؤوسه 32 وأما الكبير فبُعد التضمين 8012 وطبقاته 40 ورؤوس الانتباه 64 .

واستعملوا في تدريبها إطار Fax المعتمد على مكتبة JAX في المعالجة الرقمية المصفوفية.

وبعد التدريب ضُبطَ النموذج ضبطًا دقيقا بنصوص مولدة وكذا بشرية (بلغت 200 ألف مثال) ومترجمة وقد نُشرَ تقرير كامل على المدونة المُستعملة أيضا [13] وفي التقييم اختبروا النموذجين على بياناتٍ لم يرَهَا وعلى مهامّ الإدراك اللغوي بمدونة MMLUالمترجمة وأدَّت العربية فيها في أكبر نموذج 53.9 وأما في الترجمة فأداءه كذلك متوسطا في كل اللغات بلغ 43 على مقياس روج ROUGE (ولم يحدد الباحثون أداء العربية لوحدها)] .12[

- عائلة ArabianGPT

نماذج Arabian GPT من معمل الروبوتات وإنترنت الأشياء في جامعة الأمير سلطان تجربة جديدة متواضعة صدر منها أربع نماذج أساسية (ويُنتبه إلى أن الورقة ذكرت اثنين والاثنين الباقيان صدرا نُشرا فقط) 134 مليون و345 مليون و800 مليون و1.5 مليار معامل

الأول له نافذة سياق 768 جزءٍ لغوي والباقون بطول 1024. وكلها تشبه مواصفات AraGPT ما بين عدد الطبقات والرؤوس وأساسُهُمَا نموذج .GPT-2 استعمل مقسم sentence piece بعدة 64000 جزءا مميّزا ومعه كذلك عائلة مقسمات متنوعة الأحجام والنوع باسم .Aranizer أدى النموذج 345 مليون أداء جيدا مقارنة بنموذج BLOOM 7B في مقياس MMLU بقيمة 25.7. ضُبطت هذه النماذج على مهمة التحليل العاطفي وأظهرت نتائج مفاجئة إذ عينة صغيرة من الأمثلة أبدت تحسن أداء قارب 95% بالدقة وأما على مهام التلخيص والأسئلة والأجوبة فلم تؤدي بشكل قوي نظرًا لصغر الأحجام واستعمل في التدريب مكتبة Hugging Face مع تعديلات طفيفة[15] [34].

- عائلة علام

نماذج علام من أقوى ما صدر للغة العربية وأحدثه طورته الهيئة السعودية للبيانات والذكاء الاصطناعي (سدايا (SDAIA والمركز الوطني للذكاء الاصطناعي (SCAI) وله اصداران :

الأول كان ضبطًا لنموذج لاما-2 وله ثلاث نسخ 7 و 13 و 70 مليارَ معامل، والإصدار الثاني دُرِّبَ فيه نموذج من الصفر وحجمه 7 مليار .

عدد الأجزاء اللغوية التي دُرب عليها النموذج بالعربية بلغت 540 مليار جزء لغوي نصفها مترجم ونصفها مجموع وأما الأجزاء من الإنجليزية بلغت 1.2 ترليار. استعمل في تدريب النموذج مقسم لاما-2 وعُدِّلَ عليه بإضافة أجزاء لغوية جديدة وتضمينات هذه الأجزاء وُضعت بطريقتين إما جعلها عشوائية من البداية أو حساب تضمينها من متوسط قيم تضمينات أجزاءها .

عدد كروت الشاشة المُستعملة في التدريب بلغت 1024 كرت A100 واستُعمل إطار Megatron[35] كأساس للتدريب ثم طُوِّر وغُيِّرَ عليه .

ثم بعد تدريبه قاموا بضبط النموذج ضبطًا دقيقا على مدونة نصية بلغت 12 مليون مثالٍ ثم صُفِّيَت للنصف والذي علا تقييمه نظرًا لجودته .

ثم بعد ذلك قاموا بتنسيق (عملية التفضيل) ببناء مدونة فيها 250 ألفَ مثال ثلاثي

)أمر [جواب مرغوب][جواب مرفوض](

واستعملوا فيه خوارزمية .DPO وأخيرًا في تقييم النموذج فله طريقتان الأولى مُأَتمتة (آلية) فحصد مثلًا في MMLU ما قيمته 75.92 وفي مقياس الرياضيات 56.8، والثانية مُؤَنسنة باستعمال نظام محادثة يقارن ويُقيِّم فيها الناظر بين الأجوبة من علام وغيره من النماذج] .16[

---

[34] كاتب هذه السطور ممن يعمل على هذه النماذج وما زال التطوير والتحسين على قدم وساق نسأل الله التيسير .

[35] على الرابط: https://github.com/NVIDIA/Megatron-LM :

ملاحظة ختامية

وآخر ما نتعرض إليه في عرض النماذج المفككة هي نماذج الشركات الضخمة والتي لا علمَ لنا بها إلا اسمها وتقييماتها وهي نماذج شركة OpenAI[36] (وقد سبق في فصل تقييم الأداء بعض من قيَّمه) وشركة Google[37] وشركة Claude[38] وكلها تدعم اللغة العربية ولكن لا معلومات لدينا على البيانات التي تدربت عليها ولا حجمها ولا طُرق تدريبها لأنها مغلقة المصدر تجارية التوجه وهي للآن تُعتبر أكثر النماذج اعتمادًا عليها واستعمالا (خاصة ChatGPT) إذ مما يلاحظ من خلاصات الأوراق السابقة أن أداء النماذج العربية المدربة قد يتجاوزها في بعض المهمات بطفيف ولكن عند الاستخدام التطبيقي فالغالب يستعملها نظرًا لوجود إطار برمجي API سهل يُوفِّرُ عناء تنزيل النماذج الضخمة وكلفة تشغيلها ما بين عتاد وخبير. ولكن لا يُنكر أيضا وجود أُطُر برمجية لبعض النماذج السابقة مثل نماذج شركة كوهير [39]وعلام (ولكنَّه محدود جدا على خوادم IBM) وكذا فإن لنموذج جيس واجهة استخدام يُسْتَجرب منها .[40]وهنا إشارة إلى أن التعامل اللغوي في النماذج اللغوية غالبه (بل قل كله) متعامل مع اللغات اللاتينية (وهناك جهود أخرى في اللغة الصينية ولكنها غير مفيدة للعربية لاختلاف الأنظمة) من اليسار إلى اليمين اتجاها ولغة ولا بُدَّ من دراسة دقيقة للمقسمات الموجودة وتأثير ذلك على اللغة العربية ومحاولة تحسينها أو حتى تطوير جديدٍ ليتناسب مع طبيعة اللغة العربية، وهناك بعض الدراسات مثل [64] التي قارنت مقسمات متنوعة على مهمة التصنيف .

---

[36]وهو المعروف بتطبيقChatGPT

[37]وهو المعروف باسم GEMINI وقديما كان اسمهBARD

[38]وهو أقل النماذج شهرة مقارنة بسابقيه واسمه Sonnet 3.5 وأداؤه العربي جيد.

[39]على موقعهم/https://docs.cohere.com :

[40]على الرابط/https://jais.inceptionai.ai/c :

4.4. نموذجا المرمز والمرمز المفكك

نشير مرة أخرى لسبب إفراد هاتين الهكليتين (البُنيتين) عن السابقة: والسبب هو أن جميع النماذج التابعة لهذه الهيكلية تُعتبر صغيرة مقارنة بالنماذج العملاقة السابق الكلام عنها وأيضا لأن غالب التوجه البحثي في تطوير النماذج اللغوية يكاد ينفرد تماما بتطوير نماذج مفككة من بعد عام 2022 ومع ذلك فلا بُدَّ من ذكر النماذج العربية التابعة لهذا القسم نظرًا لفائدتها العظيمة وكثرة استعمالاتها بحثيا وتطبيقيا.

1.4.4. نموذج المرمز

أول نموذج عربية صدر بهيكلية التشفير هو نموذج ArBERT عام 2020 وقد دُرب على 2.7 مليار جزء لغوي [18] ثم صدر له اصدار جديد بأربع نسخ جديدة دُربت على 8.6 مليار جزء لغوي .[41]ثم تتابعت النماذج الجديدة فصدرت سلسلة ARBERT, MARBERTعام 2021 وقد دُرب نموذج ARBERT على 6.2 مليار جزء لغوي من العربية الفصحى وأما نسخة MARBERT فقد دربت على نصوص عامية [19] وقد صدرت لهما أيضا نسخة جديدة دُربت على 27.8 مليار جزء لغوي .[42]ومن النماذج ذات هيكلية المرمز صدر نموذج GigaBERT وهو من النماذج عديدة اللغات ومنها العربية [20]. ومما صدر أيضا نموذجا "جابر وصابر JABER "وحجمه 135 مليون معلمة و SABER حجمه 370 مليون معلمة وقد دُربا على مدونة نصية حجمها 115 جيجابايت من النصوص المُصفاة من حوالي 515 جيجابايت [21]. وهناك نموذج اسمه "قريب "QARIB دُرِّبَت منه خمسُ نُسَخٍ على مدونات نصية مختلفة الأحجام أكبرها بلغ 60 جيجا بايت] 63 .[

[41]ليس لها ورقة بحثية وإنما نُشرت فقطhttps://huggingface.co/aubmindlab/bert-base-arabertv02 :

[42]أيضا ليس للنسخة الجديدة ورقة وإنما نشرت فقطhttps://huggingface.co/UBC-NLP/ARBERTv2 :

2.4.4. نموذج المرمز المفكك (الفاك)

وهذه الهيكلية هي أول ما استعمل من نموذج المحولات في ورقتها الأصلية وأول نموذج عربية فقد صدر عام 2021 باسم AraT5وقد حقق نتائج ممتازة في 52 اختبار متنوع وقد دُرب على مدونة نصية حجمها حوالي 250 جيجابايت 180 منها فمن تغريدات تويتر و70 منها نصوص عربية فصيحة وكل جزء استُعمل في تدريب نموذج خاص (واحد للفصحى والثاني للتغريدات) ونموذج ثالث دُرب على مجموع المدونتين [22] وصدر بعد فترة نموذج جديد AraT5v2 بنافذة سياق طولها 1024 [43]وكل هذه النماذج منشورة مفتوحة المصدر [44]وأيضا فقد صدر نموذج "ترجمان" وأصله نموذج AraT5 ثم ضُبط على مهمة الترجمة وحقق نتائج ممتازة [23] وفاقت ما يجدر الإشارة له أيضا وهو عائلة نماذج nllb من فيسبوك للترجمة وعنوان الورقة "لا لغة تُتْرَك" [24] وأداءها في العربية نموذجي. وأخيرا فلدينا نموذج AraMUS صدر عام 2023 هو من أضخم النماذج في هذه الهيكلية (حجمه 13 مليار معامل) ودُرب على مدونة حجمها 530 جيجابايت وهو من أحسن النماذج الموجودة ونتائجه تفوق كل النماذج السابقة [25] وقَد قورن مع عدد كبير من النماذج المفككة، وأخيرا يجدر الإشارة لنموذج عربي مختص بالمحادثة مع الصور (بأن تعطيه صورة ثم تسأله وتدردش معه عنها) وهو نموذج "طاوس "Peacock وهو نموذج مبني على التركيب والتأليف بدلا من التدريب من الصفر مع معيرة مقتصرة على أجزاء من النماذج و أطراف زائدة] 65[

واعلم أننا سنعرّجُ في فصل التطبيقات على أهم تطبيقات هذه النماذج إن شاء الله.

---

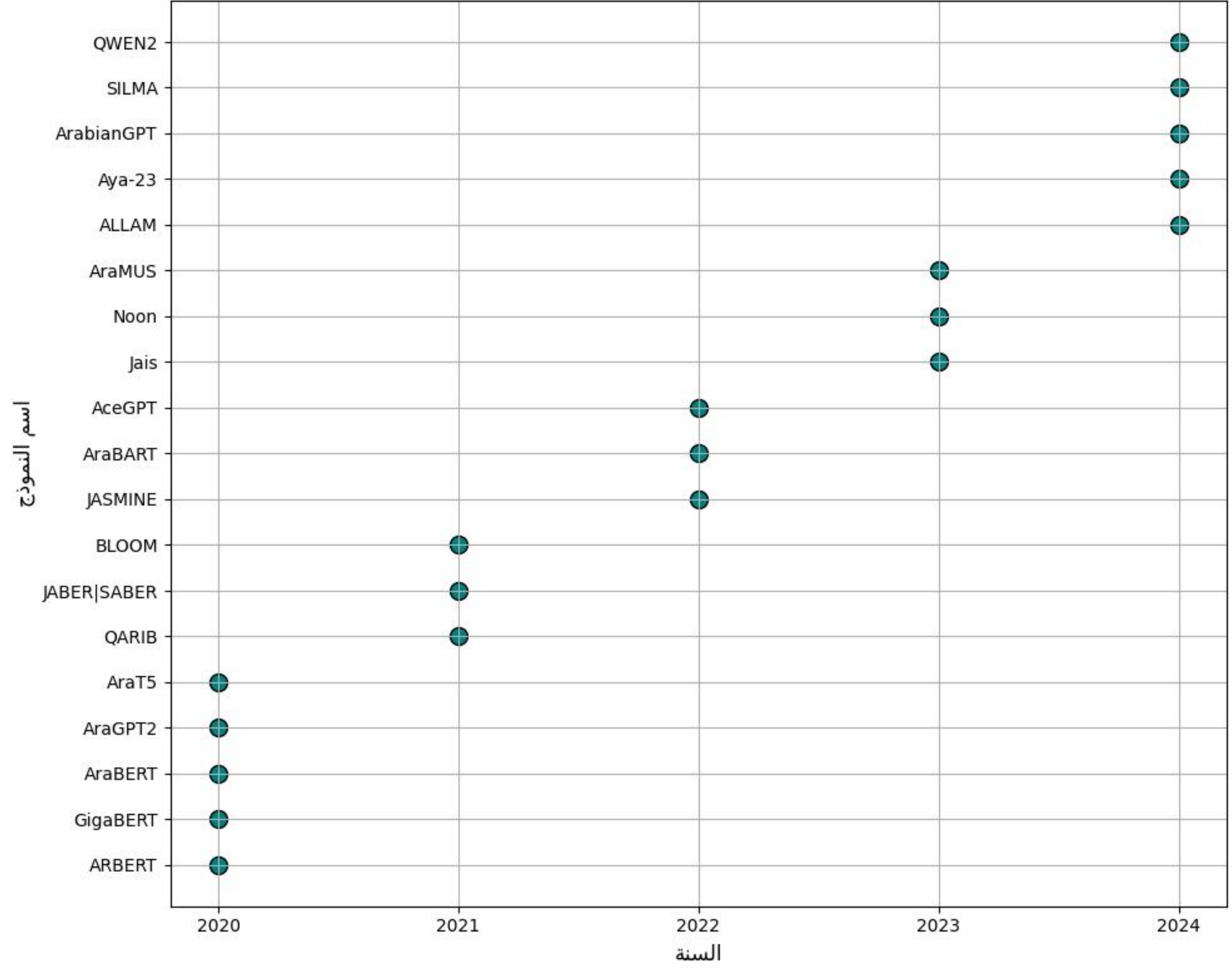
[43]لم يُصدر له ورقة وإنما نُشر فقط: https://huggingface.co/UBC-NLP/AraT5v2-base-1024

[44]الرابط: https://huggingface.co/UBC-NLP

5.4. رسوم بيانية عن النماذج العربية

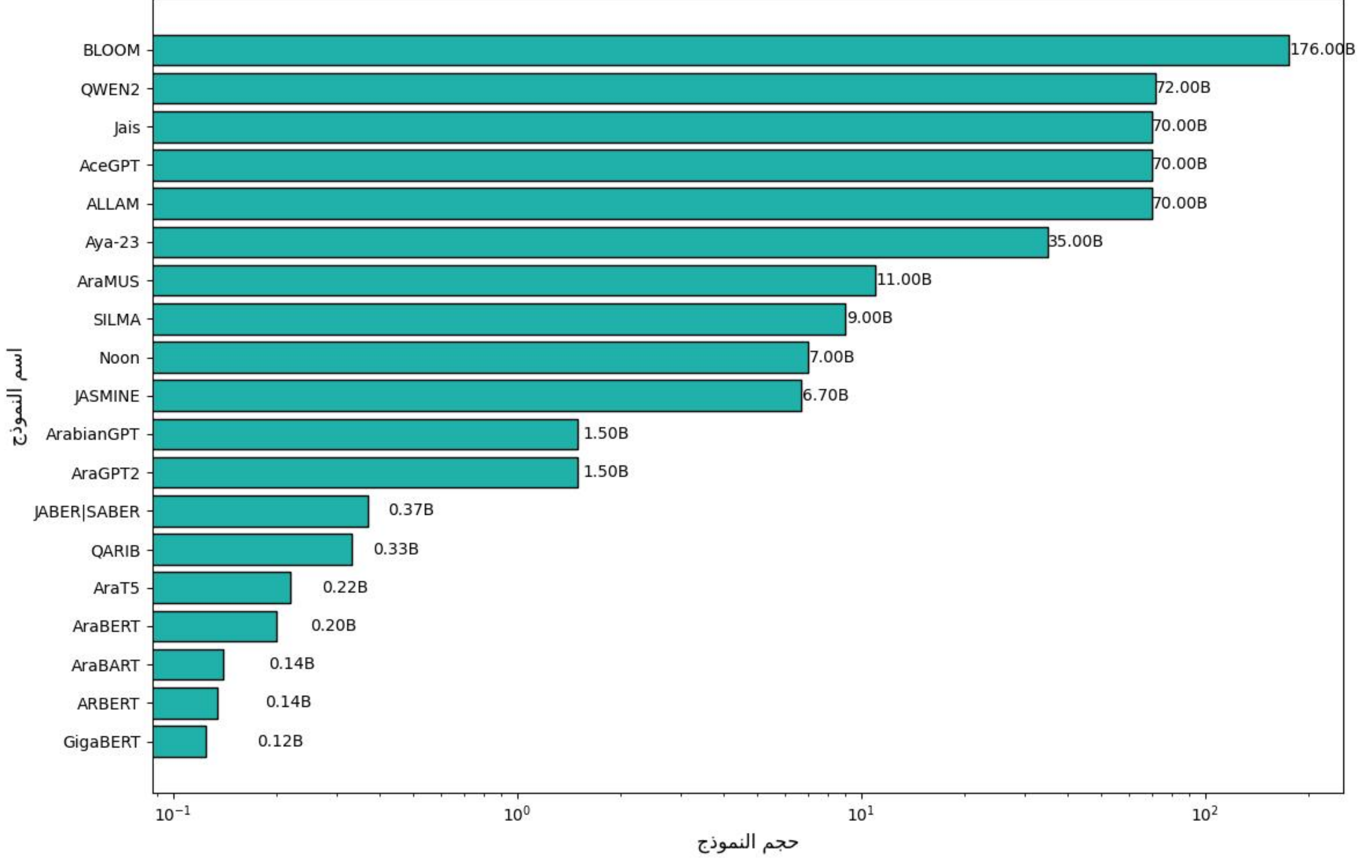
نستعرض هنا رسمين بيانيَّن أولهما في الترتيب التاريخي

شكل1: عرض تاريخي مختصر للنماذج اللغوية مع سنة إصدارها (استُعملت النسخة الأساسية للنموذج)



وأما الرسم الثاني فمتعلق بأحجام النماذج

شكل2: رسم بياني لعرض أحجام النماذج مع اسم كل نموذج منها



***ملاحظة:* تم تحجيم الرسم البياني باستعمال دالة اللوغاريتم لجعل النماذج الصغيرة أوضح وقد تم استعمال أكبر حجم من كُلِّ عائلة .**

5. الباب الرابع فـــي تطبيقات النماذج اللغوية

1.5. تمهيد

في هذا الباب نعرض أشهر التطبيقات التي تطورت أو ظهرت مع ظهور النماذج اللغوية الكبيرة. إشارة مهمة أن هذه التطبيقات كلها يُمكن استخدام النماذج الضخمة المنشورة من قبل الشركات الكبيرة ChatGPT و GEMINI و Claude فهذه النماذج الكبيرة كلها استعملت سواء بواجهة أو بالإطار البرمجي API في هذه التطبيقات التي سنسردها تاليا وقد أشير في فصل التقييم ومعاييرها إلى الأوراق البحثية التي قيَّمت بعضًا منها وفي سردنا التالية سنذكر التطبيق ثُمَّ نشير إلى أحدث وأهم التحديثات التي استجدَّت مع اللغة العربية مما يفيد المستخدم العادي أو المطور أو الباحث إذ كلٌّ منهم يُمكنه استعماله بطريقة مخصصة .

2.5. أهم التطبيقات والاستخدامات

◿ الترجمة

لا شك أن الترجمة من أهم المهمات المتعلقة باللغة العربية إذ فيها نقل المعرفة وخاصة مع قلة المصادر العلمية والمعرفية المكتوبة باللغة العربية فكان لا بُدَّ من تطوير هذا الجانب فصدرَت نماذج كبيرة متخصصة في الترجمة كما مرَّ في "ترجمان" [23] وأيضا فهناك محاولات كثيرة لتقييم هذه النماذج في مهمة الترجمة كما مرَّ في فصل التقييم ومقاييسها وكذا رأينا أن الترجمة استُعملت في توسيع المدونات المُستعملة في تدريب النماذج اللغوية العربية كما في جيس [10] وعلام [16]. وأيضا مما كَثُر فيه البحث الآن (للأسف) ترجمة اللهجات والفصحى بعضهما من بعض ومن ذلك ما كان في مؤتمر اللغة العربية الثاني في تايلاند [27]. وكذلك فقد بدأت دراسات باللغة العربية تشير إلى هذا المجال [ع1] بأنه ما زال واعدًا ويحتاج تعديلات وإصلاحات كثيرة.

◿ الأسئلة والأجوبة (الإجابة)

هذه من أهم وأشهر تطبيقات النماذج اللغوية خاصة ذات الواجهات مثل ChatGPT وتشمل أمورًا كثيرة كطلب شرح مفهوم معين أو كتابة خطة لدراسة أو زيارة أو اقتراح مصادر أو أماكن. ولكن يعيب هذا التطبيق هلوسته (بأن تذكر وتكتب جوابا غير حقيقي أو غير صحيح) النماذج أحيانا ومن الأمور كذلك أن يُطلب من النموذج تيسير (تبسيط) نص أو مفهوم معين حتى يَسهل فهمه ومن الأوراق في ذلك] 26 .[

◿ التلخيص

وهذه أيضا من أشهر تطبيقات النماذج اللغوية بحيثُ يُعطى النموذج نصا طويلا ويُطلب منه تلخيصه لآخر أصغر ويستخدمه بعض الطلبة في تلخيص الدروس والموظفين في تلخيص التقارير وغير ذلك.

◿ مهمات التصنيف وتشمل المشاعر وكشف المغلوطات (مثل الأخبار) وتصنيف الأنواع (مثل أنواع المقالات أو اللهجات)

مهمات التصنيف كثيرة وأشهرها تصنيف المشاعر بأن يُعطى النموذج نصا معينا ليصنِّفه إلى أحد المشاعر الأساسية (عادة تكون سلبي أو إيجابي أو طبيعي محايد) ومن ذلك [28] وكذلك من مهام التصنيف تصنيف النصوص إلى أنواعها الأساسية

كتصنيف المقالات الإخبارية وكذلك تصنيف النصوص المغلوطة (الكاذبة). ومن الأمور الجديدة وهي تصنيف النصوص حسب جودتها لاستعمالها في تدريب النماذج اللغوية وكذا تصنيف النصوص حسب لهجتها.[45]

- التعرف على الكيانات المُسمَّاةNER

وهذه من أقدم المهمات في المجال وهي في التعرف على الكيانات المُسماة وهي فئات محددة مسبقًا مثل الأسماء، والمنظمات أو المواقع أو الرموز الطبية أو التعبيرات الزمنية أو الكميات أو القيم النقدية أو النسب المئوية إلخ ومن آخر ما صدر ما كان في مؤتمر العربية الثاني] 29.[

- التشكيل النصي

وهذه مهمة من خصائص العربية وتهدف إلى "وضع علامات التشكيل الصحيحة على النصوص الخام" ومن أحسن النماذج الجديدة الصادرة نموذجا CATT (نموذج مرمز ونموذج مرمز مفكك) فقد أظهرا أداء عاليًا في التشكيل] 30.[

- تصحيح الأخطاء النصية واللغوية

تصحيح الأخطاء يُقصد به إصلاح النص المعطى (الحاوي عددا من الأخطاء سواءً كانت لغوي أو كتابية) وإرجاع نص صحيح وهذه الأخطاء إما أن تكون حقيقية أو موضوعة (مُصَنَّعة) لتدريب النموذج ومن الأوراق في هذا المجال مما استعمل ChatGPTفي التصحيح [31] وبعضهم قام بتدريب وضبط نماذج محددة] 32[

- استرداد المعلومات

وهذا المجال مما بزغ نجمه في المجال حديثا بما يُسمى "التوليد المُسترجِع RAG "بحيث ساعدت النماذج الجديدة على تشكيل تضمينات ذات جودة عالية ساعدت في فرز النصوص في متجهات محفوظة في قواعد معينة واسترجاع معلومات معينة حسب السؤال الموجه إلى النموذج وهنا مباحث تطوير نماذج عربية تنتج متجهات ذات جودة عالية لأنها مستخدمة في عدد كبير من التطبيقات. ومن الأوراق البحثية التي قامت بدراسة طُرق الاسترجاع مع العربية [33]. وكذا من المهام التي بُدءَ العمل عليها من السنة الماضية 2023 مهمة القاموس المعكوس بأن يُعطى التعريف ثُمَّ يُسترجع الكلمة المُعرفة (وهذا عكس عمل القاموس) وهذه ورقة المركز الأول في مؤتمر العربية الثاني] 34 .[

- كتابة النصوص البرمجية

وهذه أيضا مما ساعدت فيه النماذج أي كتابة النصوص البرمجية لعدد كبير من اللغات البرمجية خاصة بايثون واللغات البرمجية المشهورة ولكن في هذا التطبيق مبحث ضِعف القدرات المنطقية والبرمجية عند الطلاب وهذا مجال بحثي منفصل متعلق بمناهج وطرق التعليم وتغيّرها مع التطور البحثي.

- توليد النصوص المنوعة

هذا تطبيق واستخدام جديد ظهر مع النماذج اللغوية وهو توليد النصوص المتنوعة لتدريب النماذج اللغوية الكبيرة الأخرى وقد استعملته عدد من الشركات منها ميتا وجوجل ولكنَّ له أثرًا رجعيا سلبيا على المدى البعيد حسب بعض الأوراق البحثية [39,40]. وفي العربية فقد استُعملت هذه الطريقة في توليد عدد من مدونات النماذج السابق ذكرها إما للإختبار أو التدريب أو التقييم.

---

[45] وهناك أيضا مهمة لقياس فصاحة النص ولكن مُخرجها رقم حقيقي

3.5. ملاحظة ختامية

وهناك أيضا بعض المجالات التي خُصِّصَت بالذكر وهي المجال الطبي والقانوني والشرعي وهذه أفردت ببعض الأوراق مما سبق ذكره في قسم التقييم من تقييم النماذج قانونيا وكذا في المجال الطبي [36] وأما في المجال الشرعي فهناك الكثير من الأبحاث المتعلقة بالقرآن فمنها مافي مؤتمر العربية عام 2023 من مهمة الأسئلة والأجوبة عن القرآن الكريم [37] وكذلك هناك دراسة دكتوراه مطولة لا تخلو من فوائد عن تطبيقات الذكاء الاصطناعي في القضاء [ع2] ودراسة ماجستير عن "توظيف تقنيات الذكاء الاصطناعي في الدعوة إلى الله" [ع3] ودراسة أخرى عن "توظيف الذكاء الاصطناعي في خدمة السنة النبوية" [ع4] وهناك بحوث عربية كثيرة جدا بدأت تظهر في الفترة الأخيرة فيما يتعلق بالذكاء الاصطناعي أخلاقيا وتعليميا وترجميا [ع5]. ومن المجالات أيضا التي لم تشتهر ولكن الجهود فيها كبيرة مجال تخصيص النماذج في مجال معين (إما بالضبط الدقيق أو بالتدريب من الصفر) كالبنوك والشركات عموما، خاصة في مهمات استخراج المعلومات وتحويلها من الهيئة غير المُهيكلة إلى صيغة مُهيكلة سهلٌ التعامل معها (كالجداول) وأيضا مهمات التلخيص كما في مجال التداول والتقارير المالية (توليدا أو استفادة).

وأشير ختاما إلى مجال ظهر مع النماذج اللغوية وهو "هندسة الأوامر" وخلاصته في كتابة الأوامر المناسبة التي تحصل بها على أحسن الردود من النماذج وله دراسات كثيرة ومما صدر عن العربية هذه الورقة] 35.[

ولكثرة التطبيقات والأبحاث في خدمة العربية فإنه يصعب حصرها بل يعسر الإشارة لكل جهد فيها فجدير بالذكر مجاميع الأوراق البحثية التي تصدر عن المؤتمرات الخاصة باللغة العربية مثل] 38.[

6. الخاتمة

1.6. الخلاصة

قدَّمنا في الفصل نظرةً شاملة مختصرة عن غالب ما يتعلق بالجهود المتتابعة في تطوير النماذج اللغوية الكبيرة بدءًا بمقدمة تاريخية يرمق فيها القارئ أهم المحطات التاريخية في هذا المجال وصولا لطفرة الشبكات العصبية والتي منها نموذج المحولات وما بُني عليه من نماذج كثيرة منها النماذج العربية التي بدأت من عام 2020 تقريبا ووصولا إلى سنة 2024 (ويُلاحظ أن هناك جهودا متزامنة بين الشركات والجهات البحثية وهذا يدل على أهمية النماذج للطرفين). وقد مهدنا قبل الكلام عن النماذج وتفاصيلها إلى بعض المصطلحات وتقدمات مهمة عن معايير التقييم وأشهر المدونات النصية العربية المُستعملة فيه وكذا أهم مقاييس التقييم. ثم أحصينا النماذج اللغوية العربية الضخمة المفككة وتكلمنا على كل منها باختصار فيما يتعلق بمدونة التدريب ومعالجتها وحجم النموذج ومقسمه وطُرق تقييمه. ومررنا بعدها على أهم النماذج العربية من هيكلية المرمز وهيكلية المرمز المفكك وذكرنا أشهرها. ثم عَرَّجنا ختامًا على أهم تطبيقات النماذج اللغوية سواء من تقليدية أو ما ظهر حديثا مع النماذج الجديدة.

2.6. التحديات

أما تحديات هذا المجال عربيًّا فلعله يُجمع في أصول: الأول شُحّ المصادر النصية وضُعف طُرُق توليدها (سواءً ترجمة أو استخراجا).

والثاني قِلَّة القدرات الحوسبية (سواء عتادًا أو خبرة) بشكل عام وبطء تطورها مقارنة مع الجهات الأخرى .

والثالث قلة النتاج عربي اللسان فيما يتعلق بمعالجة اللغة العربية فهذا يُوسع الفجوة بين الطالب والمجال إلا بتعلمه لغةً أخرى، وأيضا هذا يجعل غالب النتاج العلمي مفيدًا للجهات الأجنبية أكثر منه للعرب.

3.6. الإقتراحات

فاقتراحنا أن ينظر الباحثون ما بين أيديهم من قدرات وموارد ثم يبنوا عليه مشروعًا بحثيا لا أن يتبعوا التيار كيفما كان ثمَّ يعلقوا فيه بلا نتاج ذا بال وأيضًا ينبغي الاعتناء بالعربية الفُصحى خالصة لأن هذا المجال يُمكن أن يساعد على إعادة الناس للغتهم الصحيحة ولو نسبيا باستعمالهم إياها في أنظمة المحادثة بدلا من الاهتمام بالعامية وتهشيش ما بقي عند من الفصحى وتقوية ما عندهم من العامية، وأخيرًا فيجب أن نستفيد مما يستجد تقنيا ولكن لا ينبغي أن نحصر العمل والاستفادة فقط بالجديد فإن هناك مجالات كثيرة لغوية (سواءً حوسبيا أو نظريًّا) ينبغي العناية بها وعدم اهمالها.

4.6. البحوث المستقبلية

أما الأبحاث المستقبلية فنقترح زيادة البحث في أنظمة الترجمة خاصة وكل ما يساعد على توسيع المدونة العربية قبل بذل الجهود في تدريب النماذج اللغوية لأن كل الأبحاث دلت وما زالت تشير إلى أن جودة المدونات واتساعها هو أهم أمر في إنشاء نموذج جيد. وأيضا فينبغي أن تكون هذه المنشورات بذرةً لتشكيل قاعدة بحثية عربية تنشر بالعربية ما يتعلق ويفيدها وخاصة لتسهيل وصول الناس إلى المعلومة ولإنعاش الشبكة البحثية العربية.

7. المراجع العربية

[ع1] صونيا أسمهان حليمي. "الترجمة الآلية وأنموذج الترجمة إلى العربية: المشهد الراهن." مجلَّة الإيسيسكو للُّغة العربيَّة 1 .313-295 :(2024) no. 1 ,

[ع2] الجلعود، أروى. أحكام تطبيقات الذكاء الاصطناعي في القضاء. جمعية قضاء، 2024. (أصلها رسالة دكتوراه)

[ع3] الحربي، ابتسام بنت عبد الله. توظيف تقنيات الذكاء الاصطناعي في الدعوة إلى الله. المعهد العالي للدعوة والاحتساب، جامعة الإمام محمد بن سعود الإسلامية، 2019. (رسالة ماجستير)

[ع4] آل مُعَدِّي, محمد بن عبد الله and ,محمد بن عبد الله. "توظيف الذكاء الاصطناعي في خدمة السنة النبوية-رؤية في أبرز المخاطر والتحديات." الدراية 182.24-161 :(2024) no. 24 ,

[ع5] المركز العربي الديمقراطي في برلين له عدد من البحوث والكتب المنشورة الجديدة في المجال وكذا هناك عدد كبير من دراسات الدكتوراه والماجستير وقد أشرنا لبعضها وهناك عدد من الكتب المنشورة مثل كتب الدكتور "طعيمة" وكذا ما ينشره المجمع من دراسات لسانية وكتب (ككتب مركز الملك عبد الله قديما) وكذا منشورات الجامعات مثل سعود والإمام ولعل هذا انطلاقة لحصر الجهود العربية في الذكاء الاصطناعي وترتيبها

[ع6] الهيئة السعودية للبيانات والذكاء الاصطناعي. معجم البيانات والذكاء الاصطناعي. د. ن, 2022.

[ع7] شركة صخر ومعلوماتها الأساسية على الموسوعة الحرة Wikipedia
https://ar.wikipedia.org/wiki/%D8%B5%D8%AE%D8%B1_(%D8%B4%D8%B1%D9%83%D8%A9)

[ع8] ملخص محاضرات في اللسانيات التطبيقية للدكتور مصطفى طويل (منشور على الشبكة)
https://moodle.univ-chlef.dz/ar/course/info.php?id=1426

8. المراجع الأجنبية

[1] Antoun, Wissam, Fady Baly, and Hazem Hajj. "AraGPT2: Pre-trained transformer for Arabic language generation." arXiv preprint arXiv:2012.15520 (2020).

[2] Workshop, BigScience, Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow et al. "Bloom: A 176b-parameter open-access multilingual language model." arXiv preprint arXiv:2211.05100 (2022).

[3] Laurençon, Hugo, Lucile Saulnier, Thomas Wang, Christopher Akiki, Albert Villanova del Moral, Teven Le Scao, Leandro Von Werra et al. "The bigscience roots corpus: A 1.6 tb composite multilingual dataset." Advances in Neural Information Processing Systems 35 (2022): 31809-31826.

[4] Dubey, Abhimanyu, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur et al. "The llama 3 herd of models." arXiv preprint arXiv:2407.21783 (2024).

[5] Abdul-Mageed, Muhammad, Abdelrahim Elmadany, Alcides Inciarte, and Md Tawkat Islam Khondaker. "JASMINE: Arabic GPT Models for Few-Shot Learning." In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pp. 16721-16744. 2023.

[6] Huang, Huang, Fei Yu, Jianqing Zhu, Xuening Sun, Hao Cheng, Dingjie Song, Zhihong Chen et al. "AceGPT, Localizing Large Language Models in Arabic." arXiv preprint arXiv:2309.12053 (2023).

[7] Zhu, Jianqing, Huang Huang, Zhihang Lin, Juhao Liang, Zhengyang Tang, Khalid Almubarak, Abdulmohsen Alharthi et al. "Second Language (Arabic) Acquisition of LLMs via Progressive Vocabulary Expansion."

[8] Seelawi, Haitham, Ibraheem Tuffaha, Mahmoud Gzawi, Wael Farhan, Bashar Talafha, Riham Badawi, Zyad Sober, Oday Al-Dweik, Abed Alhakim Freihat, and Hussein Al-Natsheh. "ALUE: Arabic language understanding evaluation." In Proceedings of the Sixth Arabic Natural Language Processing Workshop, pp. 173-184. 2021.

[9] Hellsten, Simon. "Incremental Re-tokenization in BPE-trained SentencePiece Models." (2024).

[10] Sengupta, Neha, Sunil Kumar Sahu, Bokang Jia, Satheesh Katipomu, Haonan Li, Fajri Koto, William Marshall et al. "Jais and jais-chat: Arabic-centric foundation and instruction-tuned open generative large language models." arXiv preprint arXiv:2308.16149 (2023).

[11] Hugging face page about new Jais models (include the new report) https://huggingface.co/inceptionai/jais-adapted-70b (has info about all)

[12] Aryabumi, Viraat, John Dang, Dwarak Talupuru, Saurabh Dash, David Cairuz, Hangyu Lin, Bharat Venkitesh et al. "Aya 23: Open weight releases to further multilingual progress." arXiv preprint arXiv:2405.15032 (2024).

[13] Singh, Shivalika, Freddie Vargus, Daniel Dsouza, Börje F. Karlsson, Abinaya Mahendiran, Wei-Yin Ko, Herumb Shandilya et al. "Aya dataset: An open-access collection for multilingual instruction tuning." arXiv preprint arXiv:2402.06619 (2024).

[14] Üstün, Ahmet, Viraat Aryabumi, Zheng-Xin Yong, Wei-Yin Ko, Daniel D'souza, Gbemileke Onilude, Neel Bhandari et al. "Aya model: An instruction finetuned open-access multilingual language model." arXiv preprint arXiv:2402.07827 (2024).

[15] Koubaa, Anis, Adel Ammar, Lahouari Ghouti, Omer Necar, and Serry Sibaee. "ArabianGPT: Native Arabic GPT-based Large Language Model." (2024).

[16] Saiful Bari, M., Yazeed Alnumay, Norah A. Alzahrani, Nouf M. Alotaibi, Hisham A. Alyahya, Sultan AlRashed, Faisal A. Mirza et al. "ALLaM: Large Language Models for Arabic and English." arXiv e-prints (2024): arXiv-2407.

[17] Yang, An, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li et al. "Qwen2 technical report." arXiv preprint arXiv:2407.10671 (2024).

[18] Antoun, Wissam, Fady Baly, and Hazem Hajj. "Arabert: Transformer-based model for arabic language understanding." arXiv preprint arXiv:2003.00104 (2020).

[19] Abdul-Mageed, Muhammad, AbdelRahim Elmadany, and El Moatez Billah Nagoudi. "ARBERT & MARBERT: Deep bidirectional transformers for Arabic." arXiv preprint arXiv:2101.01785 (2020).

[20] Lan, Wuwei, Yang Chen, Wei Xu, and Alan Ritter. "An empirical study of pre-trained transformers for Arabic information extraction." arXiv preprint arXiv:2004.14519 (2020).

[21] Ghaddar, Abbas, Yimeng Wu, Ahmad Rashid, Khalil Bibi, Mehdi Rezagholizadeh, Chao Xing, Yasheng Wang et al. "Jaber and saber: Junior and senior arabic bert." arXiv preprint arXiv:2112.04329 (2021).

[22] Nagoudi, El Moatez Billah, AbdelRahim Elmadany, and Muhammad Abdul-Mageed. "AraT5: Text-to-text transformers for Arabic language generation." arXiv preprint arXiv:2109.12068 (2021).

[23] Nagoudi, El Moatez Billah, AbdelRahim Elmadany, and Muhammad Abdul-Mageed. "TURJUMAN: A public toolkit for neural Arabic machine translation." arXiv preprint arXiv:2206.03933 (2022).

[24] Costa-jussà, Marta R., James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi et al. "No language left behind: Scaling human-centered machine translation." arXiv preprint arXiv:2207.04672 (2022).

[25] Alghamdi, Asaad, Xinyu Duan, Wei Jiang, Zhenhai Wang, Yimeng Wu, Qingrong Xia, Zhefeng Wang et al. "Aramus: Pushing the limits of data and model scale for arabic natural language processing." arXiv preprint arXiv:2306.06800 (2023).

[26] SalahEldin, Yousef, and Caroline Sabty. "Simplify: Automatic Arabic Sentence Simplification using Word Embeddings." In Proceedings of ArabicNLP 2023, pp. 418-422. 2023.

[27] Abdul-Mageed, Muhammad, Amr Keleg, AbdelRahim Elmadany, Chiyu Zhang, Injy Hamed, Walid Magdy, Houda Bouamor, and Nizar Habash. "NADI 2024: The Fifth Nuanced Arabic Dialect Identification Shared Task." arXiv preprint arXiv:2407.04910 (2024).

[28] Al-Thubaity, Abdulmohsen, Sakhar Alkhereyf, Hanan Murayshid, Nouf Alshalawi, Maha Omirah, Raghad Alateeq, Rawabi Almutairi, Razan Alsuwailem, Manal Alhassoun, and Imaan Alkhanen. "Evaluating ChatGPT and bard AI on Arabic sentiment analysis." In Proceedings of ArabicNLP 2023, pp. 335-349. 2023.

[29] Jarrar, Mustafa, Nagham Hamad, Mohammed Khalilia, Bashar Talafha, AbdelRahim Elmadany, and Muhammad Abdul-Mageed. "WojoodNER 2024: The Second Arabic Named Entity Recognition Shared Task." arXiv preprint arXiv:2407.09936 (2024).

[30] Alasmary, Faris, Orjuwan Zaafarani, and Ahmad Ghannam. "CATT: Character-based Arabic Tashkeel Transformer." arXiv preprint arXiv:2407.03236 (2024).

[31] Kwon, Sang Yun, Gagan Bhatia, El Moatez Billah Nagoudi, and Muhammad Abdul-Mageed. "Beyond English: Evaluating LLMs for Arabic Grammatical Error Correction." arXiv preprint arXiv:2312.08400 (2023).

[32] Salhab, Mahmoud, and Faisal Abu-Khzam. "Araspell: A deep learning approach for arabic spelling correction." arXiv preprint arXiv:2405.06981 (2024).

[33] El-Beltagy, Samhaa R., and Mohamed A. Abdallah. "Exploring Retrieval Augmented Generation in Arabic." arXiv preprint arXiv:2408.07425 (2024).

[34] Sibaee, Serry, Abdullah Alharbi, Samar Ahmad, Omer Nacar, Anis Koubaa, and Lahouari Ghouti. "ASOS at KSAA-CAD 2024: One Embedding is All You Need for Your Dictionary." In Proceedings of The Second Arabic Natural Language Processing Conference, pp. 697-703. 2024.

[35] El-Sheikh, Abdelrahman, Ahmed Elmogtaba, Kareem Darwish, Muhammad Elmallah, Ashraf Elneima, and Hassan Sawaf. "Creating Arabic LLM Prompts at Scale." arXiv preprint arXiv:2408.05882 (2024).

[36] ALMutairi, Mariam, Lulwah AlKulaib, Melike Aktas, Sara Alsalamah, and Chang-Tien Lu. "Synthetic Arabic Medical Dialogues Using Advanced Multi-Agent LLM Techniques." In Proceedings of The Second Arabic Natural Language Processing Conference, pp. 11-26. 2024.

[37] Malhas, Rana, Watheq Mansour, and Tamer Elsayed. "Qur'an QA 2023 Shared Task: Overview of Passage Retrieval and Reading Comprehension Tasks over the Holy Qur'an." Proceedings of ArabicNLP 2023 (2023): 690-701.

[38] مجموع الأوراق البحثية في مؤتمري اللغة العربية الأول في سنغافورة والثاني في تايلند مما يحسن مراجع لكثرة أوراقه وتطبيقات وأفكاره وكذا من المواقع المفيدة arxiv ففيه تنشر مسودات مشاريع كثيرة فابحث فقد عن Arabicفي صندوق البحث واختر العناوين.

- Habash, Nizar, Houda Bouamor, Ramy Eskander, Nadi Tomeh, Ibrahim Abu Farha, Ahmed Abdelali, Samia Touileb et al. "Proceedings of The Second Arabic Natural Language Processing Conference." In Proceedings of The Second Arabic Natural Language Processing Conference. 2024.
- Sawaf, Hassan, Samhaa R. El-Beltagy, Wajdi Zaghouani, Walid Magdy, Ahmed Abdelali, Nadi Tomeh, Ibrahim Abu Farha et al. "Proceedings of ArabicNLP 2023." In Proceedings of ArabicNLP 2023. 2023.

[39] Shumailov, Ilia, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross Anderson, and Yarin Gal. "AI models collapse when trained on recursively generated data." Nature 631, no. 8022 (2024): 755-759.

[40] Long, Lin, Rui Wang, Ruixuan Xiao, Junbo Zhao, Xiao Ding, Gang Chen, and Haobo Wang. "On llms-driven synthetic data generation, curation, and evaluation: A survey." arXiv preprint arXiv:2406.15126 (2024).

[41] Chang, Yupeng, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen et al. "A survey on evaluation of large language models." ACM Transactions on Intelligent Systems and Technology 15, no. 3 (2024): 1-45.

[42] Hu, Taojun, and Xiao-Hua Zhou. "Unveiling LLM Evaluation Focused on Metrics: Challenges and Solutions." arXiv preprint arXiv:2404.09135 (2024).

[43] Khondaker, Md Tawkat Islam, Abdul Waheed, El Moatez Billah Nagoudi, and Muhammad Abdul-Mageed. "Gptaraeval: A comprehensive evaluation of chatgpt on Arabic nlp." arXiv preprint arXiv:2305.14976 (2023).

[44] Alyafeai, Zaid, Maged S. Alshaibani, Badr AlKhamissi, Hamzah Luqman, Ebrahim Alareqi, and Ali Fadel. "Taqyim: Evaluating Arabic nlp tasks using chatgpt models." arXiv preprint arXiv:2306.16322 (2023).

[45] Alghamdi, Emad A., Reem I. Masoud, Deema Alnuhait, Afnan Y. Alomairi, Ahmed Ashraf, and Mohamed Zaytoon. "AraTrust: An Evaluation of Trustworthiness for LLMs in Arabic." arXiv preprint arXiv:2403.09017 (2024).

[46] Abdelali, Ahmed, Hamdy Mubarak, Shammur Absar Chowdhury, Maram Hasanain, Basel Mousi, Sabri Boughorbel, Yassine El Kheir et al. "LAraBench: Benchmarking Arabic AI with Large Language Models." arXiv preprint arXiv:2305.14982 (2023).

[47] Almazrouei, Ebtesam, Ruxandra Cojocaru, Michele Baldo, Quentin Malartic, Hamza Alobeidli, Daniele Mazzotta, Guilherme Penedo et al. "AlGhafa Evaluation Benchmark for Arabic Language Models." In Proceedings of ArabicNLP 2023, pp. 244-275. 2023.

[48] Elmadany, AbdelRahim, El Moatez Billah Nagoudi, and Muhammad Abdul-Mageed. "ORCA: A challenging benchmark for Arabic language understanding." arXiv preprint arXiv:2212.10758 (2022).

[49] Seelawi, Haitham, Ibraheem Tuffaha, Mahmoud Gzawi, Wael Farhan, Bashar Talafha, Riham Badawi, Zyad Sober, Oday Al-Dweik, Abed Alhakim Freihat, and Hussein Al-Natsheh. "ALUE: Arabic language understanding evaluation." In Proceedings of the Sixth Arabic Natural Language Processing Workshop, pp. 173-184. 2021.

[50] Elmadany, Abdelrahim, Ahmed El-Shangiti, and Muhammad Abdul-Mageed. "Dolphin: A Challenging and Diverse Benchmark for Arabic NLG." In Findings of the Association for Computational Linguistics: EMNLP 2023, pp. 1404-1422. 2023.

[51] Atef, Adel, Bassam Mattar, Sandra Sherif, Eman Elrefai, and Marwan Torki. "AQAD: 17,000+ arabic questions for machine comprehension of text." In 2020 IEEE/ACS 17th International Conference on Computer Systems and Applications (AICCSA), pp. 1-6. IEEE, 2020.

[52] Abdallah, Abdelrahman, Mahmoud Kasem, Mahmoud Abdalla, Mohamed Mahmoud, Mohamed Elkasaby, Yasser Elbendary, and Adam Jatowt. "Arabicaqa: A comprehensive dataset for arabic question answering." In Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 2049-2059. 2024.

[53] لا يوجد لمدونة بلسم ورقة بحثية وإنما نُشر على موقعهم بضع معلومات عنها ومنها جوانب الاختبار وقد رفعوا أمثلة تطوير من كل مهمة https://benchmarks.ksaa.gov.sa/b/balsam

[54] Koto, Fajri, Haonan Li, Sara Shatnawi, Jad Doughman, Abdelrahman Boda Sadallah, Aisha Alraeesi, Khalid Almubarak et al. "Arabicmmlu: Assessing massive multitask language understanding in arabic." arXiv preprint arXiv:2402.12840 (2024).

[55] Alghamdi, Reem, Zhenwen Liang, and Xiangliang Zhang. "Armath: a dataset for solving arabic math word problems." (2022).

[56] Hijazi, Faris, Somayah AlHarbi, Abdulaziz AlHussein, Harethah Abu Shairah, Reem AlZahrani, Hebah AlShamlan, Omar Knio, and George Turkiyyah. "ArabLegalEval: A Multitask Benchmark for Assessing Arabic Legal Knowledge in Large Language Models." arXiv preprint arXiv:2408.07983 (2024).

[57] Naous, Tarek, Michael J. Ryan, Alan Ritter, and Wei Xu. "Having beer after prayer? measuring cultural bias in large language models." arXiv preprint arXiv:2305.14456 (2023).

[58] Siddhant, Aditya, Junjie Hu, Melvin Johnson, Orhan Firat, and Sebastian Ruder. "Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalization." In Proceedings of the International Conference on Machine Learning 2020, pp. 4411-4421. 2020.

[59] Soliman, Abu Bakr, Kareem Eissa, and Samhaa R. El-Beltagy. "Aravec: A set of arabic word embedding models for use in arabic nlp." Procedia Computer Science 117 (2017): 256-265.

[60] Sabri, Tarik, Said Bahassine, Omar El Beggar, and Mohamed Kissi. "An improved Arabic text classification method using word embedding." International Journal of Electrical & Computer Engineering (2088-8708) 14, no. 1 (2024).

[61] Lo, Andrew W., and Manish Singh. "From ELIZA to ChatGPT: The Evolution of NLP and Financial Applications." (2023).

[62] Darwish, Kareem, Nizar Habash, Mourad Abbas, Hend Al-Khalifa, Huseein T. Al-Natsheh, Houda Bouamor, Karim Bouzoubaa et al. "A panoramic survey of natural language processing in the Arab world." Communications of the ACM 64, no. 4 (2021): 72-81.

[63] Abdelali, Ahmed, Sabit Hassan, Hamdy Mubarak, Kareem Darwish, and Younes Samih. "Pre-training bert on arabic tweets: Practical considerations." arXiv preprint arXiv:2102.10684 (2021).

[64] Alyafeai, Zaid, Maged S. Al-shaibani, Mustafa Ghaleb, and Irfan Ahmad. "Evaluating various tokenizers for Arabic text classification." Neural Processing Letters 55, no. 3 (2023): 2911-2933.

[65] Alwajih, Fakhraddin, El Moatez Billah Nagoudi, Gagan Bhatia, Abdelrahman Mohamed, and Muhammad Abdul-Mageed. "Peacock: A Family of Arabic Multimodal Large Language Models and Benchmarks." arXiv preprint arXiv:2403.01031 (2024).

[66] Nacar, Omer, and Anis Koubaa. "Enhancing Semantic Similarity Understanding in Arabic NLP with Nested Embedding Learning." arXiv preprint arXiv:2407.21139 (2024).